DOCUMENT RESUME

ED 182 319 TM 009 973

AUTHOR Brickell, Henry M.; Aslanian, Carol B.
TITLE Field Testing Educational Products. Aid Number 3 in Instructional Aids Series.
INSTITUTION Western Michigan Univ., Kalamazoo. Evaluation Center.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Sep 79
NOTE 130p.; For related document, see ED 031 954

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Curriculum Development; Curriculum Evaluation; Data Collection; *Evaluation Methods; Evaluators; *Field Studies; *Instructional Materials; Program Administration; Research Design; Summative Evaluation; *Teaching Procedures; Technical Writing; *Textbook Evaluation

ABSTRACT

Based on the evaluators' experience with a large-scale curriculum development project, this manual is designed to aid those who plan and implement field tests of educational products (instructional procedures and materials) in natural school settings. Within the procedures discussed, problems likely to arise are described. The steps in implementing field tests are: (1) general considerations, including alternatives, purposes, criteria, and planning; (2) management, including financing, negotiations, scheduling, staffing, evaluation staff training, community relations, and monitoring; (3) design, including site selection, sampling, data, instruments, data collection, data processing, and data analysis; and (4) reporting (both during field testing and after field testing), including content, format, timing, and distribution. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED182319

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Field Testing Educational Products

September, 1979

EVALUATION CENTER

COLLEGE OF EDUCATION
WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

TM009 973

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Mary Anne Bunda

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Aid Number 3
in
Instructional Aids Series

WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

INSTRUCTIONAL AIDS SERIES

The Instructional Aids Series is published by the Evaluation Center, Western Michigan University. It provides materials that are useful to those who teach evaluation theories and practices and to those who are engaged in independent studies of evaluation.

Additional copies may be obtained at a nominal fee to cover reproduction and mailing costs.

For information about other publications in this series, see the inside of the back cover of this document.

All correspondence in reference to the Instructional Aids Series should be addressed to:

Mary Anne Bunda
Editor, Instructional Aids Series
The Evaluation Center
Western Michigan University
Kalamazoo, Michigan 49008

EVALUATION CENTER

Daniel L. Stufflebeam
Director

Mary Anne Bunda

James R. Sanders

ABOUT THE CENTER

The Evaluation Center is committed to advancing the science and practice of evaluation. To serve its purpose the Center has developed components in research, development, instruction, dissemination, and service. Over the past ten years the staff of the Center has had significant experience in these areas working with school districts, state departments of education, foundations, and federal agencies.

The *Research Component* is designed to develop and test evaluation theory. Particular thrusts are testing the effects of school-based evaluation offices on school practice, and determining the special evaluation skill and knowledge profiles required by various groups in education. The *Development Component* is designed to work on ways and means of implementing evaluation theory. Included are techniques for meeting special evaluation objectives, new data collection instruments, and special procedures for communicating evaluative findings and relating them to practical problems. The *Instructional Component* is designed to provide in-service training in evaluation to various local, state, and national groups. While the Center is not a credit-granting agency, it collaborates with instructional units throughout Western Michigan University for the development and implementation of pre-service evaluation curricula. Emphasis in this work is at the doctors and masters levels. The Center serves as an instructional laboratory wherein students majoring and minoring in evaluation can gain practical evaluation experience. The Center is also committed to the development of novel and needed evaluation training materials and strategies. The *Dissemination Component* is designed to communicate research findings, new ideas, and materials to the education profession. *The Service Component* is designed to assist agencies in meeting their evaluation needs through two primary services. These are the design and implementation of evaluation studies and the development and installation of agency evaluation systems.

The staff of the Center has been involved in conceptualizing the CIPP Evaluation Model, researching the Item Sampling technique, developing the Advocate Team Technique, developing over 100 standardized tests, designing and conducting many in-service evaluation programs in evaluation. Other accomplishments include authoring the PDK book, EDUCATIONAL EVALUATION AND DECISION MAKING, and numerous journal articles and papers, designing and conducting many evaluation projects, and installing evaluation systems in schools, government agencies, and research agencies. Overall, the Center staff has directed externally funded projects in evaluation totaling more than four million dollars.

Field Testing EDUCATIONAL PRODUCTS

by
Henry M. Brickell
and
Carol B. Aslanian

Policy Studies in Education

*This document was previously published by
Policy Studies in Education, a division of
the Academy for Educational Development

September, 1979

Aid No. 3 in Instructional Aids Series

The work presented and reported herein was performed pursuant to a contract from the National Institute of Education, Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Note

CDC (short for "Cognitive Development Curriculum") is a fictitious acronym adopted for the **actual** curriculum development project which supplies the illustrations used throughout this manual.

RDO (short for "Research and Development Organization") is a fictitious acronym adopted for the **actual** research and development organization which created CDC.

TABLE OF CONTENTS

Acknowledgements	ii
Introduction	1
CDC and RDO	1
Development of CDC Units	2
Field Trial of CDC Units	3
Problems in the CDC Field Trial	4
I. General Considerations	5
Alternatives	5
Inspection	5
Simulation	10
Pilot Trial	11
Pilot Test	12
Field Trial	13
Purposes	17
Criteria	18
Planning	22
II. Management	24
Financing	25
Cost Computation	25
Cost Sharing	27
Revenue Sources	30
Negotiations	32
Scheduling	35
Staffing	39
Evaluation Staff Training	40
Administrative Staff	40
Specialized Staff	44
Field Staff	45
School Staff Training	47
Local Administrators	48
Specialized Staff	49
Product-Using Staff	51
Community Relations	53
Monitoring	55
Activities to be Monitored	55
Monitoring Techniques	58

III. Design 60
Site Selection 61
Defining the Universe and Population 61
Number of Sites 62
Random Sampling 63
Sample Descriptors 64

Sampling 66
Complete Census Versus Sampling 66
Forms of Sampling 67
Size of Sample 70
Unit of Sampling 72
Attrition 74
Drawing the Sample 75

Data 80
Site Characteristics 80
Conditions of Use 81
Attitudes Toward the Product 81
Effect on School Program 82
Pupil Performance 82

Instruments 84
Review Purposes of Field Test 84
Review Types of Data to be Collected 85
Re-Examine Product to be Tested 85
List Criteria 87
Re-Examine Data Processing Capabilities 88
Re-Examine Reporting Requirements 89
Review Literature and Existing Instruments 89
Select Data Sources 90
Specify Plan for the Instruments 90
Develop Items for Instruments 91
Assemble Items into Instruments 91
Testing the Instruments 92

Data Collection 93
Administering Instruments 94
Collecting and Returning Data 97

Data Processing 98
Methods 99
Location 101
Timing 102

Data Analysis 104

IV. Reporting 108
During Field Testing 108

After Field Testing 109

Audiences 109

Content 111
Progress Reports 111
Product Reports 112

Format 115

Timing 116

Distribution 117

Acknowledgments

We are pleased to acknowledge an earlier work which has aided the writing of this one.

The Southwest Educational Development Laboratory (SEDL) in Austin, Texas – one of the regional educational laboratories formerly supported by the USOE National Center for Educational Research and Development and now supported in part by the National Institute of Education – published a significant but little-noted book on field test procedures. George M. Higginson and the staff of the SEDL Prototype Program Installation System Development Branch compiled ideas and techniques to help their Branch appraise in natural school settings the products being developed by four SEDL instructional programs: Bi-Lingual Education, Mathematics Education, Multicultural Social Education, and Early Childhood Education. As Higginson and the staff explained in the Preface to their work, much correspondence, research, sorting, and conceptual realignment went into the production of their book – activities necessary because of the scarcity of information about field test procedures. Higginson and his staff called the resulting publication *Calipers*.

Calipers is a good book. It is not perfect, but it is a long way from blank paper. The table of contents lists the many concerns of a field tester. Because it was written in the infancy of research and development in education, it is strong on definitions, including a 65-item glossary. It devotes a fair amount of space to concepts related to field testing, such as research, development, and dissemination. Because it was written at a point in time when SEDL was just beginning to turn out products requiring field testing, *Calipers* is longer on the principles than on the practices of field testing. Moreover, it approaches

the topic from a systems design viewpoint and is strong on flow charts and PERT diagrams.

This manual goes beyond *Calipers*. It offers principles but it also recommends specific practices. Coming at a slightly more mature point in the life history of research and development in education, it deals more clearly with field testing as such, assuming that the reader is already familiar with related research and development concepts. It assumes, as well, that the reader has a basic vocabulary for labelling those concepts. Since it is not based on library research but on field experience, it does not contain a bibliography. The authors have unquestionably been influenced by their professional reading of evaluation literature over the years, but there are no particular works – other than their own earlier writing – which supplied direct content for this manual. That is perhaps its strength and its weakness.

The initial manuscript was submitted to three organizations for staff critiques: 1) The Evaluation Center at Western Michigan University, 2) The Research and Development Center for Cognitive Learning at the University of Wisconsin, and 3) The Northwest Regional Educational Laboratory in Portland, Oregon. Staff members with varied backgrounds gave assorted suggestions for improvements, many of which have been incorporated into this final version. A number of suggestions they made were not accommodated, however, and thus this manual cannot be regarded as having their full endorsement.

Introduction

This manual is designed to help those who must plan and execute large-scale field tests of educational products in natural school settings. Experience with large-scale field testing is quite rare in education and the literature available on the subject is sparse. However, there is rising professional interest in validating the utility and the effects of instructional procedures and materials through well-designed evaluations conducted in field locations. This book is intended both to elevate professional standards and to increase professional skill.

The work draws upon the experience of the authors in observing and conducting large-scale field tests of educational products over a number of years. The authors faced the problem of drawing on that experience without inundating the reader with it. Rather than taking a scattering of examples from various projects, requiring the reader to learn the separate context of each example, this manual takes all its illustrations from the authors' experience as external evaluators of a single curriculum development project. The examples taken from that experience will give readers a vicarious understanding of what large-scale field testing can be like in actual practice and will protect them from the assumption that their own experience as field testers will ever parallel the orderly scheme laid out in this manual.

CDC and RDO. The curriculum development project, hereafter referred to as CDC – a fictitious acronym – was created by a research and development organization, hereafter referred to as RDO – a fictitious acronym – in cooperation with ten local education agencies (LEAs). RDO specializes in curriuclum development in elementary and secondary education, attracting outside funding from both private and public sectors.

CEC was a systematic design for a new instructional program. The program was intended to help students in grades K-12 achieve several thousand learning goals developed by RDO with the help of the LEAs. The goals were progressively sequenced starting with the elementary grades, moving through the junior high grades, and ending with the senior high grades.

Development of CDC Units. The development of curriculum units to help pupils achieve the goals was the major effort of the CDC. RDO in cooperation with the ten LEAs developed approximately 100 units as examples of what the program would look like in the classroom. Each unit was addressed to one or more of the CDC pupil goals. However, the 100 units were not sufficient to accomplish all of the goals established for the program and were illustrative rather than comprising a complete, sequential curriculum. Each unit was relatively brief, ranging from about five to about 40 instructional hours and averaging about 20 instructional hours. Each consisted of a teacher's guide plus an actual kit of resource materials and/or a list of resource materials needed to teach the unit.

Each unit was developed by one of the ten cooperating LEAs under the supervision of RDO and according to its guidelines, procedures, and samples. The actual unit developers were curriculum specialists and classroom teachers in the LEAs.

Each draft unit underwent both substantive and editorial reviews, first in the LEA which had created the unit and then at RDO. In some cases, reviewers in the LEAs and at RDO returned the units to the curriculum specialists and teachers who had created them, requesting modifications. After the units passed final review at RDO, they were scheduled for a field test. However, for a complex of reasons, a rigorous field test turned out to be impossible and a less rigorous "field trial" had to be substituted.

The first chapter of this manual distinguishes between a field test and a field trial and explains why it is often wise for a trial to precede a test. It also makes clear that the similarities outweigh the differences. That is why the CDC field trial experience provides serviceable illustrations for this manual on field testing. Thus, although the term "field trial" is used throughout to differentiate the informal, quasi-experimental assessment of the CDC units from the more formal, more rigorous, experimental assessment of educational products indicated by the term "field testing," the CDC experience illuminates the territory of field testing quite well.

Field Trial of CDC Units. A total of 25 of the 100 curriculum units were completed and produced in quantity and went through the field trial which supplies the illustrations for this manual. The primary purpose of the field trial was to supply data to assist with the revision of the units. Most of the units were the product of initial development efforts and, although they had undergone both substantive and editorial review and revision both in the LEAs and at RDO, they had not yet had the benefit of revision based on actual use, even in a single classroom. Thus it was expected that each unit would be revised by RDO, perhaps extensively, based on data contained in its field trial report and on information available to RDO from other sources.

A secondary purpose of the field trial was to identify those units which were effective enough in their first trial to merit further investment by the funding agency. The agency assumed that some but not all units would prove successful enough in the field trial to merit revision, perhaps another field trial or field test, and eventual dissemination. Thus the field trial was intended to inform the funding agency's decisions as to which units were worth further attention and which should be abandoned.

Problems in the CDC Field Trial. Things did not go smoothly during the field trial. There were disagreements among the principal parties about why and whether and how the trial should be run. There were anxieties, surprises, disappointments, poor cooperation, superb cooperation, misunderstanding, missed deadlines, scapegoating, high achievement, cost overruns, credit taken, credit earned, competition, and teamwork. In short, things went the way they usually go in a large-scale field test. That is why the CDC case is particularly useful as an illustration. It is rich with reality.

The CDC case was not chosen because it illustrates the right way. It was chosen because it illustrates the real way. The examples scattered through the text of this manual violate as often as they confirm the advice of the authors. That is to say, the authors were not able to follow their own advice when confronted with the full reality of the CDC field trial. The case is spread upon the record to show the reader what his/her actual field testing experience is likely to be – even after he/she had read this manual. But the story is also told in the hope that it will help him/her do better.

I. GENERAL CONSIDERATIONS

Alternatives

Field testing is not necessarily the preferred method of product validation. Other methods, most of them less expensive, may be more suitable, especially for products not yet in fully mature form. (See Criteria section for guidelines to predict readiness for field testing.)

A number of alternatives is listed below, sequenced roughly according to rigor and cost. As a product matures, it becomes a candidate for successive stages in this sequence:

Inspection

Simulation

Pilot Trial

Pilot Test

Field Trial

Field Test

As mentioned above, each of these alternatives may be considered as alternatives to a field test if the product is not mature enough to be put through a rigorous test. They also might be considered preliminary steps to a field test. In other words, the results of each of these tests or trials will aid the evaluator in deciding whether the product is ready for field testing. A discussion of the early alternatives is given at some length, so that the kinds of information necessary to make the decision to field test can be illustrated.

Inspection

Inspection of the product can (and should) be accomplished by a wide range of both professionals and non-professionals. Listed below are a number of special-interest or

special-expertise groups who should review the product for various characteristics. Each group will review the product along slightly different dimensions which match their expertise. The combination of these multiple sources of information will provide the evaluator of the product with the information for product validation.

Professional

a. Cost Experts. Individuals familiar with school costs both for the initial installation and the continuing use of educational products can be asked to inspect new products to compare their initial and continuing costs to those of the products they would replace. Such comparisons can help the evaluator predict whether the new item -- even if it should prove effective in later testing -- would be likely to find acceptance in the school market.

> **CDC**
> The "resource kits" accompanying the teachers' guide for the CDC units were produced by the ten LEAs. RDO forwarded the kits to the CDC publisher for duplication along with the teacher's guides without first cost-inspecting the kits. The publisher sensed marketing trouble and notified RDO that the kits would cost from under $10 to over $200 per classroom. RDO then signaled the LEAs to hold kit costs down to $25 or less.
>
> But it was too late to correct the problem since many of the first 25 kits were virtually complete by that time. RDO could not afford even the 40 to 60 kits needed for the field trial of each unit. It ordered the publisher to produce partial kits and notified the LEAs to obtain missing items locally. But the LEAs could not always afford to acquire the missing parts.
>
> The result was that some kits were too expensive, incomplete, or in short supply (had to rotated from one teacher to another). As a consequence, the trial of some units was awkward or unrealistic. In other cases the trial was of little use because the unit could never be nationally disseminated in its field trial form.

b. Curriculum and Subject Field Specialists. The accuracy, up-to-dateness, intellectual importance, social significance, and overlap with existing curriculum content can be judged by qualified curriculum and subject field specialists simply by inspecting an educational product. Such judgments are an essential prelude to using the item in school classrooms.

CDC

Curriculum specialists in some LEAs and at RDO helped design or edit some units but there was no final inspection of curriculum content by highly qualified subject field specialists (in mathematics, science, social studies, etc.) at RDO prior to the field trial. One result was that some units had to undergo content modification following the field trial.

Almost 10 of the first 25 CDC units were designated for a single grade. Since these units were produced simultaneously in separate LEAs, content overlap was inevitable. Inspection by curriculum specialists could have identified and eliminated the overlap before the units were put into field trial, giving classroom teachers the choice of either 1) using the units as designed and thus repeating content with the children, or 2) not using the units as designed so as to eliminate repeated content.

c. Instructional Specialists. The clarity of instructional objectives, sequencing of content, reading level, choice of media, amount of repetition, chances for practice, reinforcement devices, lesson-embedded test items, and other instructional features of the product can be inspected by developmental psychologists and specialists in instructional technology.

CDC

One type of instructional specialist employed by RDO to inspect the units coming in from the LEAs was a group of well-qualified evaluators. They not only inspected but found it necessary to rewrite hundreds of the lesson-embedded test items in the units. This raised the quality of the items as measuring devices. However, since representative classroom teachers did not inspect those items as well, it was not until the field trial that it became apparent that teachers would not or could not use many of the test items while teaching the lessons. Teachers, particularly those in the primary grades, thought many items needlessly formal, saying that they could judge pupil learning by observation without administering the test items.

d. Administrators. The probable acceptability, utility, feasibility, administrative convenience, availability of necessary equipment, affordability, and fit with the existing school program can be judged by a panel of local school administrators.

CDC

If the LEAs or RDO had convened a group of local administrators to examine the units prior to the field trial, the administrators might have discovered that one of the CDC units could have been tried out only in one of the LEAs because the others lacked the specialized equipment required. This information would have been useful in predicting the ultimate market for the unit as first designed and might have led to changing it to permit wider use.

e. Classroom Teachers. The acceptability, ease of use, conformance to current curriculum, vocabulary, and probable effectiveness with various kinds of students can be estimated by classroom teachers before a product is put into school use.

> **CDC**
> Although classroom teachers worked on every unit produced by the LEAs, the units were not finally reviewed prior to the field trial by a panel of teachers who had nothing to do with their creation. Such a panel might have supplied more realistic estimates of teaching time required for the units, most of which took more time than the field trial schedule allowed.

f. Publishers. The attractiveness, convenience, cost, timeliness, competitiveness, and marketability of the product to a large body of users can be judged by publishers who have had experience with similar products.

> **CDC**
> RDO wisely had a well-known commercial publisher assist with editing and duplicating the CDC units prior to the field trials. However, RDO retained all final design decisions, using the company more as a printer than as a publisher and paying outright for all company services. This arrangement perhaps deprived RDO of the kind of hard-headed judgment a publisher has to make if his/her own money is at stake.

Nonprofessional

a. Legislators. One group likely to be concerned about the broad-scale introduction of a new educational program — not a specific product, such as a curriculum unit, but the larger program of which it is a part — is the legislators who sit as members of the education and finance committees likely to influence any legislative action that might be required to move a program into general use. By examining the general goals of the program and hearing a sample of the curriculum materials described, a representative group of legislators can predict in advance the fate the program is likely to meet in their state, should legislative action ever be needed to encourage or mandate the spread of the program.

b. School Boards. At some point in the history of a new educational product – probably just prior to its initial installation in the schools – local school board members are likely to have a determining voice in whether the new product will be used. A cross-section of board members from the types of communities for which the product has been designed can help the evaluator predict the reactions of local school boards.

c. Community. Within the community around the school, some groups may be pivotal in determining the fate of the new program. For example, employers and union leaders are likely to be especially interested in new developments in career education. Similarly, minority groups are likely to be interested in new social studies programs which touch upon the role of racial and cultural minorities. Advance inspection by representatives of such interest groups can assist the evaluators in predicting community acceptance.

d. Parents. Parental reaction to the new product, while it is likely to arise rather slowly, is well worth anticipating because it can influence whether the product is continued in use. A panel of parents, including representatives of minority and special interest groups, can be asked to inspect the product in advance and to give its judgment of its conformance with their values and their response to it as a proposed addition to the schools their children attend.

e. Students. Student opinion – unfortunately rarely sought in any systematic way by product developers and evaluators until wide-scale field trials have begun – can be gathered at a very early stage. Student enthusiasm or opposition during an initial inspection of a product can be a significant signal for the developers that the product is ready to move ahead or needs a return trip to the drawing board.

CDC
One LEA arranged for the inspection of each unit by community members, including minority spokespersons.

This allowed the LEA to refer to that screening when recommending the units to principals and teachers for use in the field trial.

Simulation

At the very early stage of product development, especially when the smallest components of a product such as individual lessons are being created, artificially-arranged simulations of instructional settings can be used to give the developers easy access and near-instant feedback on the clarity, readability, and convenience of product components. For example, some of the developers themselves can pose as the kind of professional or non-professional "inspectors" called for above. That is, they can role-play the views of administrators, teachers, legislators, minority group spokesmen, parents, and so on. This simulates inspection.

Alternatively, some of the developers can pose as students, letting other developers "instruct" them with lessons, graphics, slides or other preliminary materials they have developed. This simulates pilot trials.

The role players should, of course, simulate as closely as possible the behaviors of those they are imitating. The simulation can be substituted for inspection and/or for pilot trials, discussed below. What such artificial arrangements offer in convenience and quick feedback, of course, they lose in realism.

CDC
Neither the LEAs or RDO simulated the use of inspectors or students. On the other hand, most units were developed by groups of teachers and writers and went through intensive editing at the LEAs and/or at RDO. These arrangements served as partial substitutes for a planned simulation.

Pilot Trial

The term "pilot trial" is used here to designate an initial, small-scale use of the product under conditions not so rigorous as those indicated by the term "pilot test." Especially at the very early stages of product development, there is simply no substitute for the actual use of the materials with a sample of their ultimate targets. Even a single student, moving through the material under careful over-the-shoulder observation by the developer, followed by thorough debriefing with the developer, can be enormously informative. Similarly, products intended for use with a group of students can be given an initial trial with such a group. The cost/benefit ratio of such very early, small-scale pilot trials is usually favorable; indeed, it is rarely matched by any later evaluative steps. That is, the developer probably learns more at less cost through pilot trials than through any other technique.

The developer can act as teacher, using the initial materials with one or more classes of students. The problem with this method is that the developer usually brings forward and unknowingly uses many assumptions as to how the product is to be employed. The only way to tell whether these are being communicated to the strangers who must use the material is to have a stranger use them while the developer observes.

The developer can hand the very early segments of a product to a classroom teacher who had nothing to do with their creation, take an observer's seat in the rear of the classroom, and watch with mixed emotions as the teacher tries to use the segments according to the instructions the developer has written. Of course, the developer should let the materials speak for themselves, avoiding the temptation to give oral explanations and advice as to how the material is to be taught. If the developer wants a still more

chastening experience with an embryonic product, he/she can use the same technique with several teachers in several classes, guarding against the possibility that the first teacher had some special ability or disability which could mislead the developer when he/she is drawing conclusions about how his/her product performed.

CDC

Some of the LEAs had time to put their very earliest CDC units through a pilot trial in a single classroom. In one case, for example, the developer and the classroom teacher alternated in teaching and observing the lessons, with both writing suggestions as to how the lessons could be improved. This simple, low-cost use led to significant changes in the unit. Afterward, the developer expressed genuine surprise at how badly she had communicated her intentions in the initial draft of the teachers' guide.

But time soon became short and all LEAs abandoned pilot trials. Units were shipped to RDO without having been taught even once in a single classroom, leaving the RDO editorial boards with much clean-up and re-writing to do. RDO was not able to put those edited, re-written versions into a pilot trial. This ultimately meant that the field trial turned up many problems — some small and some large — that might have been caught and eliminated earlier.

Pilot Test

A pilot test can be thought of as similar to a pilot trial, except for the degree of rigor involved. Since rigor costs something to achieve, pilot tests will be more expensive than pilot trials.

These features, among others, can be thought of as differentiating pilot tests from pilot trials:

a. At least one full classroom of students should be used.
b. The developers should not teach the lessons.
c. The information given to the teacher should be limited to what a teacher would normally receive in using the product under field conditions.
d. The teaching circumstances should approximate normality in size of class, time allowed for lessons, equipment and materials available, amount of testing, and so on.
e. Informal observation should be supplemented by formal data-collection instruments.
f. Student learning should be tested rather than simply estimated by students and teachers.

CDC

Neither the LEAs or RDO was abln to arrange genuine pilot tests of the units. Thus, as indicated above, the units went into the field trial with flaws that could have been eliminated beforehand.

One year after the field trial concluded, RDO reported that it had revised the 25 units, reducing their number to 15. It said that it had found four principal difficulties with the units:

1. An undesirably "directive and inflexible" quality in the teachers' guides, making them rigid, time-specific, and leaving too little to a teacher's discretion.

2. Shortcomings in the subject content and intellectual substance of the units.

3. "Resource kits" accompanying the curriculum units that were too bulky and too expensive for widespread dissemination.

4. Inconsistent formats among the 25 units.

It seems clear that these are the kinds of problems that could have been eliminated through inspection, simulation, pilot trial, or pilot testing prior to the field trials.

Field Trial

The term "field trial" is intended to distinguish the activity from the more rigorous "field test," which is the main subject of this manual. The distinction between the two is roughly the same as the distinction between pilot trials and pilot tests presented above, except for the larger number of classes involved in field activities.

Because "field trials" differ in "field tests" not in purpose but in degree of scientific rigor, both terms are used in this manual to designate broad-scale uses of educational products in natural settings under conditions that hopefully approximate normal use. The question to be answered in field trials or field tests is whether the product continues to perform in the same way as it did during the pilot trials and/or pilot tests.

For products which have not yet been through some kind of pilot evaluation, field assessments are unnecessarily expensive and may waste the time of teachers and students, or possibly even miseducate students. Not until after a product has demonstrated its effectiveness in a small number of settings does it merit broad-scale use.

While pilot trials and/or pilot tests are intended to gather data to guide product revision, field assessments probably should be used to gather data to guide product installation. That is, while small-scale pilot use is an excellent way to get revision leads at low cost and with minimum inconvenience to the schools, large-scale field use is an excellent way to determine in what settings and under what conditions the product will perform as well as or better than it did in the pilot use.

Pilot uses are particularly suitable for individual segments or components of a product at the early stages of development before the instructional style and format have become crystallized and before effort and money have been poured into making a large volume of the material. For example, the time to test a new lesson format is when the first one has been produced. Long before a product is ready for field use, its basic features should have been confirmed through pilot use.

In contrast, field use is especially suitable for completely assembled products where the questions have to do with the coherence of the product as a whole, the interaction of its various parts, the total demands it makes upon the school, and so on. These questions cannot be satisfactorily answered for various types of schools except through field use.

In summary, field trials are an alternative to field tests in the same way that less-than-rigorous pilot trials may be an acceptable less-expensive substitute for pilot tests. A great

deal can be learned about a product through wide-scale field use employing informal methods of observation, relying on teacher and student opinions about the utility and effectiveness of the product, stopping short of the requirements of good experimental design in selecting the sample, allowing natural variations in the use of a product to occur without inhibition and without close observation, and so on.

More rigorous field testing is discussed in the balance of this manual.

CDC

The decision to run a large-scale field trial of the CDC units took more than a year to make. The forces shaping that decision were complex because there were at least four parties at interest: The funding agency, RDO as prime contractor and architect of the CDC units, the ten LEAs as builders of the units and prospective field trial sites, and the staff of PSE (Policy Studies in Education) as external evaluator.

The funding agency had two primary concerns: 1) to produce high quality, thoroughly validated units, and 2) to get them nationally disseminated. The CDC project was extremely expensive by most standards and the time was approaching for it to produce goods for national use.

RDO's chief interest was in bringing the units up to a point where they would perform well in a high-visibility national field test.

The LEAs' main interest was in acquiring curriculum units and using them locally in a way that would advance the cause of the program in the schools and the community.

The PSE staff's chief concern was to perform high quality evaluation services that would enhance its reputation nationwide.

The interests of the funding agency and PSE tended to coincide and ultimately prevailed. The interests of RDO and the LEAs tended to coincide and ultimately were assigned secondary importance.

At the very inception of the project, PSE laid plans to construct a "final examination" in each grade level K-12, a design that fit RDO's original plan to produce a comprehensive set of curriculum units for each grade level. Within six months, this evaluation plan was scratched as it became clear that the units would not constitute a set and would have to be evaluated singly.

Consequently, PSE laid plans for field testing the units singly in the ten participating LEAs to answer summative questions about whether the units were ready for prospective outside users. The

object was to find out for what types of students and under what kinds of conditions the units worked.

But it was already becoming clear that the LEAs could not deliver mature, pilot tested units to RDO in time for field testing. RDO was beginning to express grave doubts about the wisdom of conducting a field test.

Originally, the PSE staff had divided the unit evaluation work with the RDO internal evaluation staff: The RDO staff would accomplish formative evaluation through pilot testing; thereafter, the PSE staff would accomplish summative evaluation through field testing. But time ran out. The RDO evaluation staff (and their counterparts in the LEAs) had not been able to get the units pilot tested. Nevertheless, the staff of PSE was preparing to begin the field test. The logical solution would have been to postpone the field test for another year, allowing time for formative testing and revision of the units. But the urgency of completing at least some units for national dissemination was uppermost in the funding agency's mind. The solution was to accomplish formative and summative evaluation simultaneously.

This classic compromise solution led to the classic compromise problem: suboptimization of objectives. The resulting formative evaluation was not as insightful and helpful as RDO and the LEAs would have preferred; the resulting summative evaluation was not as rigorous as the funding agency and the PSE staff would have preferred.

As time went on, the CDC units "became less mature" while the evaluation design "became more sophisticated." That is, the units fell further below the standards envisioned for them while the emerging evaluation design made ever-higher assumptions about what the units could accomplish and about the elaborateness of the evaluation they deserved. This led to a conflict between those at RDO who were destined to revise the units and the PSE staff who were destined to measure the accomplishments of the units. RDO wanted to influence the field test design so that revision-related information would be gathered and reported. The PSE staff, responding to the funding agency's standards, wanted to push the field test design toward gathering summative data that could guide decisions about the nationwide dissemination of the units.

RDO expressed increasing concern about the appropriateness of the evaluation design for its purposes while the PSE staff expressed increasing concern about the formative obligations being placed on the field test design. RDO began to make its own plans for supplementing the work of PSE by gathering additional data to guide revision.

The final result was a field trial design that tried to satisfy all parties by collecting summative information primarily but writing field trial reports that interpreted the data so as to guide unit revision as well as nationwide dissemination. Neither objective was fully achieved.

Purposes

As indicated earlier, the purposes of field testing can be accomplished in part, at less expense and with less rigor, by field trials. Thus the purposes of field tests do not differ from the purposes of field trials.

The customary purposes of field tests are as follows:

1. To show whether the newly-developed product performs as intended under normal or near-normal use in a wide variety of conditions with populations large enough to give results that are both statistically significant and credible.

2. To show under what conditions the product does and does not perform as intended – in what kinds of socio-economic settings, with what kinds of populations, when used by what kind of teachers, with what kinds of administrative and professional support, and so on.

3. To establish whether a product works under conditions different from those in which it was developed and without the supervision of its inventors.

4. To determine the cost of installing the product.

5. To determine the amount and kinds of manpower necessary for installing the product.

6. To determine the physical facilities required for its use.

7. To determine the training requirements for the school staff.

8. To determine the amount of time necessary for the product to achieve its objectives.

9. To determine whether the product is worthy of further investment by its sponsor as well as further effort by its developers.

10. To facilitate eventual widespread acceptance of the product, if it is successful by familiarizing a cross-section of geographically-dispersed users with its existence and its characteristics.

11. To gather data on the potential commercial success of the product.

12. To guide the revision of the field tested product, the development of substitutes for it, or the creation of supplementary products.

CDC

The CDC unit field trial had the first nine of these purposes, although not all nine were accomplished. It also had the final purpose in the list above: to guide the revision of the units. This final purpose was only partially accomplished, chiefly because of its conflict with the first nine purposes. Although a combination of formative and summative purposes is common in field trials and field tests, any design which attempts to answer both formative and summative questions simultaneously has many problems.

For example, for reasons of economy and convenience as well as to make prompt use of the findings, formative test settings should be located near the original developmental site. In contrast, summative test settings should ideally be scattered nationwide in a pattern of geographic and demographic representation that will not only put the product through its paces in assorted environments but will also lend nationwide credibility to the findings. In the case of the CDC units, 50 teachers in 50 buildings in five cities scattered 3,000 miles apart constituted a very good summative sample for each unit, but an extremely inconvenient formative sample.

To take a second example, the inevitable tension between the need for funds to measure pupil learning and funds to measure teacher viewpoints was resolved in the case of the CDC units in favor of measuring pupil learning at the expense of measuring teacher viewpoints.

The effect of problems such as these was that the PSE staff was not able to produce useful information for the unit revisers. RDO met this need in its own way by paralleling the field trials with separate data collection activities. It conducted panel reviews with teachers after they had taught the units, sponsored reviews by teachers who had not taught the units, arranged critiques by the RDO editorial staff, and asked teachers to make marginal notations in duplicate copies of the teachers' guides as to how the units should be modified. In retrospect, this was an extremely wise move by the RDO to get what it needed to guide unit revision.

Criteria

There are a number of criteria that can be used in determining whether a product is ready for field testing. These have to do not only with the stage of maturity a product has reached — obviously a key criterion — but also with the cost of conducting the test relative to the cost of creating the product, the likelihood that the data collected will in fact be used, the provision of sufficient guidance for instrument development, and the existence of enough funds to carry out the work. These and other criteria appear in further detail below.

1. **The degree of development the product has reached,** as judged by its developers, outside experts, a cross-section of prospective users and the evaluator. The product must be complete, understandable, and transportable. It must work without further attention from or supervision by those who developed it.

2. **The efficiency of testing the product in its current form.** If the product is a component of a larger program, it may be wiser to delay field testing until the complete set of materials is available. This is an especially important consideration if the performance of the component would be affected by its use in conjunction with the remainder of the program.

3. **The adequacy of the inspection, simulation, or pilot use already accomplished.** As indicated above, it is unnecessarily expensive to the funding agency, time-consuming for the project staff and the evaluation staff, and burdensome to the schools to field test a product which has not yet demonstrated its usefulness and effectiveness in small-scale runs.

4. **The objectives of the product** – perhaps expressed in terms of behavioral outcomes – are clear enough to guide the creation of instruments for assessing pupil learning. Developers are often unclear as to exactly what effects their procedures and materials will actually have on their target audiences. Pilot use of the materials employing trial instruments can help determine what effects the product has and thus guide the development of the final outcome-measuring instruments.

5. **The variables which may affect the success of the product are well enough understood** to be controlled or, if allowed to vary, observed systematically. This is essential if the performance of the product or its differential performance in varied settings is to be explained.

6. **The ultimate use of the product is large enough** to warrant the time, cost, and inconvenience of field testing. A product which, even when in saturation use, will be employed in only a limited number of settings probably does not merit the effort required to conduct a formal field test. Small-scale pilot trials or pilot tests would be more suitable.

7. The degree of certainty that **the field test data will actually be used** for one of the purposes given in the previous section.

8. **Accessibility to the required kinds of school sites and the serious committment** of the staff in those sites to carrying out the field test according to the design supplied by the evaluator.

9. **The availability of a staff of evaluators** qualified to handle both the design and the mechanics of the field test, as described later in this manual.

10. **Field test results will be ready early enough** to guide the key decisions to be made about the product. Too often, decision-makers will reach a deadline before the field test results can be reported to them. Under certain circumstances, field testing is obviously ceremonial.

11. **The costs of producing and purchasing the product** are so low that it does not justify field testing. For example, large-scale, rigorous field testing of paperclips would be wasteful in that the risks of actual production and use are so low for the producers and users that they need not be reduced through formal field testing.

12. **Sufficient funds** are available to conduct all aspects of the field test at a similar level of quality. These include all the steps of planning, negotiations, staff training, instrument development, data analysis, reporting, and publication and distribution of reports.

Only if a product meets the above criteria does it merit the effort entailed in conducting a large-scale field test. The preliminary screening possible through inspection, simulation, pilot trial, pilot test, and field trial can either 1) eliminate unworthy products at an earlier, less costly stage or 2) bring them to a more finished state so that they merit field testing.

CDC

The CDC units could not meet many of the criteria for field testing. As explained earlier the units were underdeveloped; they had not been adequately inspected or piloted; they needed further revision before large-scale testing.

They failed on other criteria as well. For example, the project staffs in several of the LEAs did not have a strong commitment to carrying out a field trial according to a design supplied by the PSE staff as external evaluator (see criterion eight above). Some were unfamiliar with field trial procedures and considered them burdensome to themselves as well as objectionable to teachers.

Moreover, the PSE staff was not able to command the services of LEA evaluators and was unable to supply a qualified cadre of site monitors to do the field work required.

The result was that there was a slippage in most aspects of the test procedures and schedule.

Planning

Running a field test involves administering an extremely complex set of geographically-dispersed operations and coordinating the work of a very large number of people. It is extremely difficult to anticipate, much less to control, all events. And yet unless all operations are coordinated and move forward on an interlocked timetable, the field test cannot succeed. In short, field testing is an activity that absolutely requires detailed planning.

The major topics to which the plan itself must be addressed are those appearing as major headings in the remainder of this manual. They include financing, staff training, monitoring, data collection, reporting, and many others.

Because there are numerous parallel strands of activity in a field test, and because they tend to be interlocked and contingent upon one another, a PERT-like network planning scheme is necessary to embrace their complexity.

It is essential that the planners remain flexible and be ready to re-plan the schedule when necessary. Things often go wrong in field testing because most of the participants – those on whom success depends – are not likely to be on the payroll of the evaluator and are likely to have obligations they regard as more significant or more urgent than carrying out their field test duties. Thus they are not subject to direct control by evaluators and may not respond to indirect controls such as appeals for consideration.

Evaluators should read the remaining sections of this manual with the understanding that the activities described cannot merely be approached serially. They must be planned in advance and an appreciable number of them undertaken simultaneously. As the outline

itself will indicate, the management demands of field testing are greater than the evaluation design demands. The activity must be staffed accordingly.

CDC

The original plan for the CDC field trial called for units to be delivered to the LEAs in the summer with the field trial to begin in September. But unit development took longer than expected. Serial shipment of the units did not begin until late November and the field trial was scheduled for the following January-May for most units.

The PSE staff at first relied on RDO to monitor the field trial and the LEAs to carry it out. But when the PSE staff visited the LEAs in early April, it found that some had not begun using units they had had since February. The PSE staff realized that monitoring, guidance, and assistance had been unsatisfactory.

For example, the PSE staff had not surveyed the LEAs to find which eighth grade home economics courses had the child development content needed for one CDC unit. And it had not worked with the LEAs to identify sufficient classrooms to test the approximately 10 units designated for a single grade.

The result was that the PSE staff found it necessary to begin aggressive monitoring and energetic problem-solving in mid-April, telephoning LEAs daily to check progress, ordering surplus units shipped from one LEA to another in search of available classrooms, asking the LEAs to bring additional LEAs (not part of the original ten) to supply over 200 classrooms for surplus units, asking another LEA to arrange a statewide field trial of an industrial arts unit that most other LEAs had no equipment to teach, and in many other ways using emergency procedures to make up for inadequate planning.

II. MANAGEMENT

The management of field testing – that is, the organizational administrative, logistical, supervisory, and financial considerations – is extremely demanding. Relationships must be maintained with many parties at interest – the product developers, the field test sites, the agency funding the field test, the evaluator's own staff and field personnel, the evaluator's subcontractors, outside agencies interested in the outcomes of the field test, and so on. Cooperative planning and cooperative execution require a great deal of time. The evaluator is likely to find that he/she has taken on a major training function as well as a major publishing function before his/her work is completed. He/she will find that he/she has not only to schedule events but to monitor their occurrence and to make up somehow for shortfalls and failures.

This chapter deals with many of those problems.

Financing

Field testing is likely to be an expensive operation. Because of the number of sites and the number of people involved as well as the possibility of unexpected events, it is an activity for which advance cost estimates are difficult to make with accuracy. Thus contingency funds are highly advisable.

At the same time, the interest of many parties in conducting the field test and using the findings often makes it possible to share costs among agencies and to use multiple sources of revenue. Those topics are discussed below.

Cost Computation

The table of contents for the remainder of this manual constitutes a rough set of budget categories for field testing. The relevant ones are repeated below for convenience:

a. Planning
b. Site Selection
c. Negotiations
d. Logistics
e. Evaluation Staff Training
f. School Staff Training
g. Community Relations
h. Monitoring
i. Instrument Development
j. Sampling
k. Data Collection
l. Data Processing
m. Data Analysis
n. Report Writing
o. Report Publication and Distribution

Evaluators cannot prepare an adequate budget without taking all these cost factors into account. Activity in most of the areas requires that money be spent on communications and, in many cases, on transportation.

There are advantages to a budget-by-task to make sure that all likely costs have been covered. On the other hand, many field test activities are so parallel and interlocked with each other that more than one can be accomplished at the same time. For example, in making a trip to a potential candidate site, it is possible to make a roster of logistical requirements for additional equipment which would need to be met if that site were to be chosen. Thus, if a budget-by-task is created, it needs to be re-analyzed to consolidate expenses and get maximum economic efficiency.

The primary variable determining the cost of the field test is the degree of rigor wanted. If the evaluator is seeking a high degree of certainty in the conclusions reached, he/she will want: a) a comprehensive cross-section of sites, including some that are geographically inconvenient and expensive to reach; b) thorough training for his/her own evaluation staff, including field observers and their data gatherers; c) to make certian that the field test teachers are properly trained in the use of the product; d) to monitor field operations closely rather than depending upon reports from local personnel who are not accountable to him/her; e) to develop original instruments rather than using instruments already available; f) a large sample that can be subdivided many times without cell size falling below an acceptable threshhold; and g) elaborate data processing and thorough data analysis. In general, reducing costs means reducing the quality and/or the quantity of the data collected.

From the evaluator's point of view, the most attractive economies are not likely to be economies at all but simply ways of sharing the costs with other participating agencies.

CDC
The original field trial design used by the PSE staff required 60 classrooms distributed over the ten LEAs in a pretest-posttest design

without control groups. About eight units were being tried out in this way when the funding agency called for a more rigorous design. Over the objections of the LEAs, which did not want to test classes not using the units, PSE moved the remaining units into a pretest-posttest design requiring 40 experimental groups and 40 control groups. This move raised the cost of the field trial by 33 percent, a figure the funding agency was willing to pay for the greater rigor.

Cost Sharing

It is not neccessary for evaluators to pay all the costs of conducting a field test. The costs can often be shared – and in some cases should be shared – with the product developer and/or with the field test sites. Thus some careful advance thought about who should pay what is both worthwhile and necessary.

A distinction should be made between the ordinary, recurring **operational** costs of using an educational product and the extraordinary, non-recurring **evaluation** costs incurred in testing and educational product.

Operational costs are those normally paid by school taxpayers. They include both the start-up costs (costs of installing the product) and the continuation costs (costs of maintaining the product in use over time). Start-up expenses are usually higher because the schools must purchase equipment, acquire an initial supply of consumable and non-consumable materials, provide orientation and initial training for the school staff, acquaint the community with the program, and so on.

The **evaluation** costs associated with field testing constitute a special class of one-time-only expenses that occur during the field test but never again. These include the costs of developing, administering, scoring, analyzing, and reporting findings from data-gathering instruments employed in the field test. They also include all costs of communicating

with the field test sites, on-site monitoring during the field test, training local school personnel to administer data collection instruments that are specific to the field test rather than an integral part of the product itself, and so on.

Sharing Operational Costs. The most important feature of a field test from a cost-sharing viewpoint is the fact that the product will be tested in schools whose operational costs are already being paid by taxpayers. This represents a form of cost-sharing with local school districts, making it unnecessary for evaluators to create a complete instructional environment, which would be excessively costly. But there is another reason, for the evaluator to avoid creating a complete instructional environment - a reason even more important than economy. The purposes of field testing require that the new product be used in near-natural circumstances under ordinary field conditions.. Perhaps the most fundamental of all "ordinary field conditions" is the fact that school districts must pay for the products they install and operate. That is, most districts do not get special funding for installing and operating most programs. This means that most educational products cannot be widely disseminated unless school districts are willing to pay for them.

One customary purpose of a field test, as indicated earlier, is to gather data on the potential commercial success of the product. That is, to determine whether it is salable. If salability is a criterion, the field test should be designed so that the test schools have to pay exactly the costs that the ultimate adopter of the product would hvae to pay. Those include both the start-up costs and the continuation costs. That is, the field test should be run only in school districts willing to pay those costs.

Moreover, as explained later, one criterion for site selection is the willingness of the local school district to continue the program at district expense if it proves successful. This is

important in case the district is to be used as a demonstration site following the field test. The simplest way to identify such districts is to choose those willing to pay start-up costs. Those are the ones most likely to continue using the products. It is obvious that any product that school districts find so attractive they will purchase it in untested form is likely to be even more marketable in tested form. In short, a product for which districts will pay start-up costs gives evidence of salability.

On the other hand, if salability is not a criterion, if the only questions to be answered by the field test are how well and under what conditions the product works, the product developer and/or the evaluator may share the operational start-up costs without defeating the purpose of the field test. But even when salability is not in question, the evaluator may be able to persuade the school district or the product developer to pay all operational costs.

Evaluators can argue that the developer has far more to gain than the evaluator in seeing that the product is tested in its most attractive, most convenient form. Thus, it can be argued that the developer should want to print multiple copies, and ship them directly to the field test sites ready for field testing, thus saving time and costs for the evaluator.

Similarly, evaluators can argue that local school districts should supply the trainers to acquaint local teachers with the new product, letting the evaluator limit his/her effort to training the trainers. Teacher training is a legitimate expense for the local school district inasmuch as the local schools will continue to benefit from the skills the teachers learn.

Sharing Evaluation Costs. It will be considerably more difficult for evaluators to persuade the local school districts that they ought to pay for such evaluation costs as devel-

oping instruments, collecting data, and reporting results. While the local district will benefit from the findings, it will recognize that many other districts will benefit as well. Indeed, it may believe it could learn enough about the product by letting other districts test it and avoiding the inconvenience and risk of testing it in local schools.

Evaluators may have more success in convincing the product developer to pay some or all of the evaluation costs, particularly if the developer has much to gain in reputation or in sales if the product performs well in the field test.

While the evaluator should be aggressive in seeking cost-sharing for operational costs and even for evaluation costs, he/she cannot abandon financial responsibilities for operations that touch on the quality of the field test. He/she must see that the products are delivered on time to the sites, that the local staff is properly trained to use the product and/or administer the instruments, and so on. If evaluators cannot arrange for such costs to be shared, they must budget funds for doing it themselves.

> **CDC**
> RDO as product developer shared the field trial costs by printing and shipping multiple copies of the CDC units in sufficient quantity for the field trial.
>
> The 10 LEAs as field trial sites shared costs by managing all the local logistics of receiving, logging in, distributing, collecting, and shipping the field trial instrument packages to the PSE staff.
>
> These arrangements represented tens of thousands of dollars in cost-sharing, distributing field trial costs over RDO, the LEAs, and the PSE staff.

Revenue Sources

Although there are a number of potential revenue sources -- local school districts, state education agencies, various federal agencies, philanthropic foundations, research and

development centers and regional educational laboratories, publishers, and so on, it generally proves to be the case that the agency most interested in field testing the product is the one which has paid for its development and wants to know its performance characteristics. What this means to the evaluator is that his/her revenue sources tend to be limited to those who have enough interest in the product to pay for its assessment. This almost always means the developers or their funding agency.

The only effective way for evaluators to draw on the funds of agencies which did not finance the development of the product is to have them pay either the developers or the participating local school districts for their cost-shared portion of the work. By using the discretion available to them under the grants and contracts they have received, developers as well as local school districts can ordinarily find justification for sharing the field test costs if they actually wish to do so.

CDC

Inasmuch as the funding agency was willing to pay the full costs of the CDC field trial — indeed, its insistance on certain features in the evaluation design raised considerably the price it had to pay — it was not necessary to seek other revenue sources. The probability of finding supplementary funds — except for the RDO and the LEA cost-sharing described earlier — would have been very near zero for this project.

Negotiations

Once the prospective field test sites have been drawn by the sampling procedure (see **Site Selection** later in this manual) the evaluator must approach the administrative head in each of the sites to elicit his/her interest in having his/her school district participate in the field test. Evaluators must acquaint the prospective site with both the advantages and the obligations of participating.

The advantages he/she can point out to the district include these:

1. The recognition and prestige accruing to the distict because it is taking part in an innovative program.
2. Leadership experience for the administrative staff in acting to install a new product.
3. Professional growth for the teaching staff in becoming familiar with a new movement in the profession.
4. Possible improvement in the local school program.
5. Early access to new procedures and materials that may not become available through other means for one or two years longer, even if successfully field tested.

As to the obligations the district would incur by agreeing to take part in the field test, evaluators must specify (and may need to persuade the chief administrator to supply) the full set of circumstances required. These are likely to include at least the following:

1. A cooperative spirit in meeting the field test requirements, which may prove to be less convenient and more burdensome than either the evaluator or the administrator suspects during their initial encounter.
2. The administrative time necessary to plan, coordinate, and execute the field test obligations.
3. The instructional personnel necessary to meet the sample size quota set for the site.

4. The time that any specialized support personnel -- either professional or nonprofessional -- require to assist classroom teachers in using the new product. This would include the support of such personnel as librarians and counselors, paraprofessionals, bus drivers, and so on.
5. The training time necessary for administrative leaders, classroom teachers, and support personnel to learn their roles in the use of the new product.
6. The physical facilities plus any specialized equipment needed in conjunction with product use.
7. Modifications in the school scheduling as necessary to accommodate the new procedures and materials.

The evaluator should negotiate on the principle that he/she can and should supply, as a maximum, only those resources necessary to enable the district to carry out the field test. If the evaluator is testing the important proposition that the product is attractive enough to cause a school district to pay the normal operating costs of acquiring and maintaining it in the absence of any special incentives, then he/she should offer to pay the costs of the field test procedures but should not agree to pay the costs of the instructional procedures. That is, his/her intention should be to give the school district no more help than it would get if the product were being offered for installation through normal distribution channels following a successful field test. Inasmuch as the evaluator would not supply instructional extras during the eventual distribution of the product, he/she should not do so during the field test itself.

School districts which will not agree to the terms set by the evaluators should be eliminated as candidate field test sites and replaced with random selections from the list of other eligible sites drawn during sample selection. The evaluator should make careful note of the conditions a district refused or was unable to meet and should include this

information in his/her final report, since it introduces a sampling bias. The information can be used not only to help explain product performance but also to help predict product acceptability in certain kinds of school districts during general distribution following field testing.

CDC

The PSE staff made the error of not negotiating field trial arrangements with the LEAs. The assumption that the LEAs as project participants and as authors of the CDC units would understand and execute the field trial requirements turned out to be wrong. (The PSE staff had learned too much from the competent way the LEAs had administered a major attitude survey earlier.)

The LEAs should have been told the advantages and obligations of the CDC field trials. The LEA project staffs should have been told t the advantages and obligations of the CDC field trials. The LEA project staffs should have been told in advance that they would have to survey local facilities to see whether they could accept certain CDC units requiring special equipment. They should have been told what would be in the field trial instrument package, what classroom teachers would be expected to do and what information they would be expected to supply. Particular attention should have been given to the pretest/ posttest experimental group/control group design, an esoteric arrangement which the districts had never experienced. They should have committed themselves in advance to supply the required quota of classrooms for the field trials.

The PSE staff made the mistake of assuming that all this would occur automatically, rather than taking the responsibility for making it happen. Many problems arose as a direct result.

In contrast, the LEAs handled the negotiations in recruiting additional LEA extraordinarily well. LEA staffs drew up contracts explaining exactly what would be supplied to the new LEAs – how many CDC units for how many classrooms on what grade levels in what subjects on what timetable and what the new LEAs would have to supply in return. The contract spelled out who would pay for what and covered such items as who would retain ownership of the CDC units once the field trial was concluded.

The result was a remarkably trouble-free operation in getting the additional 200 participating classrooms – in many ways the smoothest operation in the entire field trial.

Of course, the new LEAs were sites that wanted to use the CDC units and were eager to cooperate – more eager than some of the original ten LEAs. Neotiations go easily with such sites. While it can be argued that such sites are so extraordinary that the evaluator cannot generalize from the field trial results, it can also be argued that such sites are the most likely purchasers of the finished product and certainly ought to be included.

Scheduling

As indicated earlier, a field test is a complex, multi-faceted, interlocking set of activities that requires carefully-detailed scheduling – combined with a readiness to change the schedule promptly when the unexpected happens, as it usually does.

The process involves identifying the activities which must take place, building a sub-schedule for each one, combining all sub-schedules into a master activity schedule, and searching for efficiencies and concomitant economies that can be realized by conducting two or more activities in the same place at the same time with the same personnel.

Local schools differ in the season, month, week, day, and time of day at which a given activity can be carried out. These differences must be taken into account in building a schedule.

Adequate lead time is extremely important in dealing with local school districts. Some local decisions are made months in advance of execution and cannot be changed to meet later field test requirements, no matter what the attitude of local school officials. Thus the evaluator has to give ample advance notice if he/she wants his/her requirements respected. There is, incidentally, absolutely no substitute for an evaluator's understanding of the normal "rhythm" of the academic year. He/she has to understand, for example, that nothing can happen in the schools before September 15 at the earliest, or between December 15 and January 15, or after May 15 at the very latest. To take another example, he/she has to understand that October is a much longer, more optimistic, more cooperative month than April.

The operational steps in building a schedule include the following:

1. Plan backward, beginning with the due date of the final field test report. Start by allowing a reasonable amount of time for printing and delivery of the final report document, then double it.
2. Allow ample time for writing the final report, taking into account the likelihood that all interested audiences cannot be served by a single version and that, sooner or later, the evaluator will be asked to prepare alternate versions for different uses.
3. Allow time for data processing and data analysis, assuming the usual computer breakdowns, program debugging problems, erroneous runs, and re-analysis when the first analyses fail to show enough that is either significant or interesting. Allow time to make hypotheses and test them after the data have been analyzed initially.
4. Get a copy of the current school calendar for each field test site and make a combined display. The critical path for field test administration will run through the timeline for the least competent school district rather than through the longest local timeline.
5. Allow time for the data collection instruments to be returned to headquarters and allow time for logging the receipt of those materials and follow-up time to retrieve missing data.
6. Allow time for teachers to complete all data collection instruments, remembering that some will never complete the entire set. (Here is a case where over-sampling is essential to allow for the inevitable attrition and to permit data processing to begin without waiting for the last teacher to complete the forms.)
7. Allow the instructional time needed to field test the product in the classrooms. The time should be as specified in the product – no more and no less if reasonably well controlled conditions are wanted – but vacation days, teacher illnesses, assembly programs, fire drills, field trips, snow storms, strikes, and other natural and manmade

events have to be scheduled into the calendar time required to supply instructional time.

8. Allow time for teachers to be oriented and trained to the use of the new product. Both the occasions and the amount of training time will differ district by district.
9. Allow time for the field test sites to receive, open, log, process, distribute, and review the field test materials before actually placing them in the hands of classroom teachers.
10. Allow time for the product and the accompanying field test packets to be transported from evaluation headquarters to local sites. If camera-ready copy rather than multiple copies is supplied to the sites, allow additional time for local reproduction of the materials. Also allow time for supplementary shipments to replace lost or missing materials when reported by the local sites.
11. Allow time to acquaint administrators and teachers with field test procedures. Their initial acquaintance with the field test must take place long before the field test is scheduled to begin so that they can adjust their own administrative and teaching schedules as necessary.
12. Schedule each activity forming a part of the field test.
13. Combine the time schedules for individual activities into one comprehensive master activity chart.
14. Arrange for regular feedback from personnel in each activity area and adjust the time schedule realistically to accommodate to actual events.

CDC

The original field trial schedule was laid out before any units had been completed. The field trial schedule assumed that the development schedule would be met, when it was not, the field trial scheduele had to be changed, an adjustment that became increasingly difficult as the trial year grew to an end.

In August of the field trial year, the unit development schedule called for 66 out of 100 units to be ready for field trials in the subsequent school year. The schedule called for three units to be

shipped to the LEAs each week, beginning in November.

By January, only four of the 18 scheduled units had been delivered to the PSE staff and the LEAs. (The PSE staff had to acquire each unit as it came off the presses, construct a pretest and a posttest, and print and deliver the tests to the LEAs before the trial of the unit began).

By February, Eight of the 30 scheduled units had been delivered.

About the first week in March, the unit development schedule (and the quota of units to go into field trial) was adjusted downward from 66 to 49 based on deliveries up to that date.

By the second week in March, 15 of the 42 scheduled units had been shipped to the LEAs.

By the last week in March, the unit development schedule (and the field trial quota) was again adjusted downward from 49 to 25 based on deliveries up to that date.

On or about the last week in April, the 25th and final unit was delivered to the PSE staff and to the LEAs.

The field trial schedule – a schedule that included test development, printing, shipping, teacher training, pretest administration, classroom instruction, posttest administration, completion of questionnaires by teachers, shipping of field trial packets from the LEAs to the PSE staff, instrument scoring, data processing, data analysis, and the writing, printing, and shipping of the field trial reports – had to be adjusted.

Many problems arose as a result.

Uncertainty about unit delivery made it difficult for the LEAs to sustain teacher willingness to teach the units.

The arrival of many units very late in the school year put them into direct competition with teachers' normal year-end activities, such as taking spring trips and reviewing for final examinations.

One consequence was that the ten LEAs could not absorb all the units; new LEAs had to be recruited.

The PSE staff and that of its subcontractor had to be expanded. A high-pressure working atmosphere plus night and weekend work became commonplace.

The test development process became more expensive because it was protracted and the economies originally anticipated from developing many tests simultaneously were lost.

Part-time personnel who had been recruited to work in the summer had to be replaced by personnel who could work in the fall.

The delivery of the field trial reports to the unit revisers at RDO ran months behind the original schedule.

Staffing

The field test staff must, of course, be qualified to handle all field test functions. Because field testing tends to be a phased activity with operations performed in linear fashion, certain staff members are needed throughout the work while others are needed at selected stages. Because management functions must be performed throughout the field test, administrative staff are needed permanently. In contrast, such functions as test development and data processing occur only at certain points in the timetable; thus persons who specialize in such work are needed only temporarily. Therefore, it is a sensible economy to assign management tasks to permanent personnel and specialized tasks to temporary personnel, consultants, or subcontractors.

A large organization with a permanent staff of specialists can assign them to field test when their services are required and assign them to other projects at other itmes. It will have less need for consultants and subcontractors.

A small organization will need to seek outside help. Temporary appointments extending from one month to several months in length can sometimes be made in finding persons in transition between jobs, taking recent graduates while they are seeking full-time positions, or arranging for a person to take a temporary leave of absence from another job. Consultants can be drawn from universities or consulting firms day-to-day as needed, with their time either interspersed throughout the work or scheduled for single tasks on single occasions. Subcontractors are desireable for efforts which are 1)specialized, 2) large-scale and 3) long-term. Data processing and data analysis typically meet these criteria.

Of course, it will not always be possible to add and drop personnel on a perfect timetable, exactly as the need for them appears and disappears. But the staffing principle is nonetheless important: as field testing moves through its successive stages, the need for specialized personnel shifts radically. Unless such personnel can be phased in and out as needed, field test costs will be needlessly high.

Table 1 contains an illustration of how field test functions can be allocated to various kinds of personnel. The table assigns management functions to administrative staff, as would be expected, indicates that staff specialists and consultants are interchangeable for most tasks, designates some tasks for field staff and others for consultants, and singles out report writing as a specialized function.

> **CDC**
> The PSE staff shifted permanent personnel into the project and out of it on many occasions as circumstances required, brought in consultants to assist with certain aspects of the evaluation design, hired temporary personnel for occasional data tabulating, and subcontracted test development and data processing as functions that could be handled more economically by specialized agencies.

Evaluation Staff Training

Assuming that the field test staff is recruited after the field test plan is approved, and thus is not familiar either with the product or with the plan for field testing it, the staff will need both orientation and training.

Administrative Staff

The administrative staff needs to become thoroughly familiar with certain aspects of the field test in order to interpret them to others and to make sensible management decisions. Those aspects include at least the following:

Table 1

Illustrative Allocation of Staff Functions

Functions	Staff				
	Administrative	Specialized	Field	Consultant	Sub-Contractor
Management					
Planning					
Financing					
Negotiating					
Staffing					
Scheduling					
Evaluation Staff Training					
School Staff Training					
Community Relations					
Monitoring					
Design					
Site Selection					
Choices					
Instrument Development					
Sampling					
Data Collection					
Data Processing					
Data Analysis					
Reporting					
Writing					
Distribution					

Nature of the Product. Complete familiarity with the product itself is essential. Its objectives, the materials required, the classroom procedures needed, the out-of-classroom activities called for, the professional and non-professional support services required, the equipment needed, the participation by non-school personnel, required interrelations with other aspects of the school program – all these and more must be understood if the field test is to be planned and supervised properly.

Purposes of the Field Test. Because a field test can have one or more purposes and must be designed, conducted, and reported accordingly, it is essential for the administrative leaders to understand thoroughly what the test is intended to accomplish and for which sponsors.

Planning. The plans already made as well as the additional planning required must be communicated to the administrative staff just as soon as they begin their work. As indicated earlier, making plans and changing plans through feedback is the key administrative activity in field testing.

Negotiations. The agreements being sought or already reached with the field test sites must be known to the administrative leaders so that they can either conduct negotiations or monitor the keeping of any bargains already struck between the field test agency and the field sites.

Staff Training Requirements. The administrators must understand the nature and timetable of any training required by the evaluation headquarters staff, by the evaluation field staff, or by the participating school staff.

Monitoring. The field test administrators must be taught that field tests rarely run according to plan. Monitoring is necessary both to keep operations as close to plans as possible and to change the plans when operations diverge. They must be trained to develop procedures for monitoring the work of the evaluation field staff as well as the work of the participating school staff. Elaboration of this topic is presented at the end of this chapter.

Site Selection. If the sites have not been chosen, the administrative staff must be trained to create (or to understand, if criteria were included in the proposal for the field test) the criteria for selecting field sites as well as procedures to be used in making that selection. If the sites have already be chosen when the administrative staff is recruited, the reasons for their selection as well as the agreement negotiated with each one of them need to be explained.

Data and Instruments. Elaborating the description of data to be gathered and determining procedures for collecting the data are typical administrative staff responsibilities, although they are usually carried out by lower-echelon personnel. The administrators will need orientation and may need training in order to plan and supervise the work.

Sampling Procedures. The administrative staff needs to be familiar with the sampling design and the sampling procedures. The first will need to be explained to many parties during the field test and will probably need to be defended vigorously against alternative sampling suggestions; the second will need to be either aggressively monitored if carried out by local school personnel or, better still, actually executed by the evaluation field staff. It is extremely difficult to get school personnel to accept random sampling designs and to carry them out faithfully.

Data Processing and Data Analysis. Elaborating the data processing and analysis plan is a step that customarily follows the specification of data and the creation of instruments for collecting it. The administrative staff needs to be familiar with the interrelations of these activities. They will of course need to employ and supervise staff for making the processing and analysis plan and will need to prepare themselves for soliciting competitive bids for data processing unless the field test agency has the in-house capability.

Report Writing and Distribution. The administrative staff will need to be familiarized with the intended audiences, content, formats, timing, and distribution plans for written reports. In addition, they will need to know that plans will need to be made for non-written reports presented to various audiences as the field test proceeds.

Specialized Staff

Specialists — in contrast to the administrative staff described above and the field staff described below — are ordinarily selected because their existing qualifications equip them to design and conduct the field test. For that reason, they are more likely to *supply* training than to require training. However, there are likely to be project-specific topics with which they will need to become familiar. These include the following:

Nature of the Product. In order to elaborate the description of data to be collected and to create the necessary instruments, the specialized staff will need to gain complete familiarity with the product being tested. The objectives of the product, the materials required to use it, the out-of-classroom activities called for, the professional and non-professional support services required, the equipment needed, the anticipated participation by non-school personnel, required interrelations with other aspects of the school program — all these and more must be understood if the specialists are to know what data to collect and how to analyze it.

Purposes of the Field Test. Because a field test can have one or more purposes and must be designed, conducted, and reported accordingly, it is essential for the specialists to understand completely why the test is being conducted and which audiences are to be served by the findings.

Timetable. Staff specialists must be taught the timetable on which the project is intended to run. Their work in creating the technical products such as data collection instruments and data analysis specifications must be completed on schedule if the overall timetable is to be met.

Budget. The specialists must know how much money is available for performing the various tasks. Most budgets fall short of what could desirably be spent for any given task. That is, the budget ordinarily allows for a quality of work short of the ideal. The specialists must know what money is available so that they can govern the quality of the work accordingly. For example, if there is no money available for pilot testing the data collection instruments (a fairly common management error in budgeting for a field test) the specialists need to be told that as a part of their initial orientation. Knowing this early may give them time to readjust the work or propose that money be transferred from less important tasks.

Field Staff

Large-scale field tests ordinarily require that the evaluation staff somehow place field representatives on site.

To the extent that the field staff has administrative responsibilities, they will need training similar to that recommended above for the administrative staff. To the extent that they are engaged in such specialized functions as collecting data – as by conducting interviews with centrally-developed interview guides – they need training in those specialized functions.

The specialized functions for which a field staff is likely to be needed are these: 1) school staff training, 2) monitoring, 3) data collection. The training required for each will differ.

School Staff Training. The field staff will need to be thoroughly grounded in any content or procedures it is to transmit to the local school staff. The relation of the content and

procedures to the total field test plan, the specific rationale for the content and procedures being taught, the consequences to the field test results if local school staffs do not learn the content and procedures, the alternative content and procedures that would be acceptable and, if not, why not – all these need to be taught to the field staff in addition to the content and procedures they are to transmit. This is because they can expect to meet many challenges from local administrators as well as local trainees as they go about supplying the training. The field staff needs to be well versed in techniques for training adults and should be thoroughly familiar with the realities of local school operations. The field staff should be trained to use demonstrations (live or simulated) to supply ample illustrations, to get the trainees actively engaged in practicing what they are being taught rather than having them listen to it passively, and to allow ample time for trainees to ask questions in a non-threatening training environment. (For example, administrators of the school district probably should not be present when classroom teachers are reviewing what they have learned with the field staff.)

Monitoring. The field staff may be given the job of aggressive monitoring of local performance and rapid, honest feedback to evaluation headquarters on how things are going. The information they collect will not be all positive: local failures to execute assigned tasks, for good reasons or bad, are bound to occur. Reporting these can strain the relations of the field monitor with the local school – yet they must be reported. Field monitors need training in how to keep themselves acceptable to the localities while telling the full truth to headquarters. They also need training in sensing local attitudes and spotting local events – not necessarily those connected directly with the field test itself – which are likely to throw the field test off its planned track at some time in the future. Helping headquarters anticipate problems is an essential monitoring function.

Data Collection. The evaluation field staff may be used either 1) to attend to the mechanics of data collection or 2) to serve as a part of the instrumentation. Training for the first entails learning about the instruments to be administered; their distribution and return; on-site logging before shipment to headquarters; checking returned forms for completeness, clarity, and ease of processing; understanding local school channels for distribution and collection of forms, including administrative protocols; and so on. Training for the second entails learning the rationale for the instrumentation, the nature and intended uses of the data being collected, the nature of validity, reliability, objectivity, comparability, and related matters. For the second function, simultaneous training of the entire field staff at headquarters by the specialized staff is likely to prove less expensive than any other means and will be required to establish the objectivity of the field staff and the comparability of what they collect. Undoubtedly, their training for this function will require the use of actual or simulated situations in which they can practice their data-collection skills.

School Staff Training

The staffs of the local schools cooperating in the field test will undoubtedly need orientation and training. This is especially true for those who must actually use the products. While training in field test *purposes* and *procedures* can be as elaborate and detailed as necessary to assure good execution, training in the *purposes* and *procedures of product use* should be limited strictly to those that will ordinarily be used during the distribution of the product following the field test. This is in keeping with a principle discussed earlier. The principle is that the product should be tested under natural field conditions — the kind of conditions that will exist following the field test. In short, the evaluator must

use the kind of *product training* that the ultimate distributors (such as publishers' salesmen) will use.

Top-echelon administrators are likely to be oriented to the field test during the initial negotiations for selecting the site. But lower-echelon administrators are likely to have only a smattering of information about the field test, even after the negotiations have been completed.

Local Administrators

The administrative staffs of the participating schools will need to be oriented to the field test -- especially to what is expected of them as administrators -- long before the test gets underway. As explained above, many key school decisions are made months ahead of execution and cannot be changed except at great cost and inconvenience, if at all, once they are made. Even when a change can be accomplished, staff attitudes toward the field test can be soured as a result. Thus it is essential for the lower-echelon administrators who must deal face-to-face and day-to-day with the teachers, counselors, and librarians who are using the product to be oriented to the field test and to their responsibilities just as early as possible.

Their orientation should include at least the following topics:

1. Nature of the Product
2. Purposes of the Field Test
3. Planning
4. Negotiations

5. Staff Training Requirements
6. Sampling Procedures
7. Data Collection
8. Report Writing, Publication, and Distribution

Other local personnel may have administrative interests in the field test. For example, in local schools with a strong teachers' association or union, the association or union leader will want to learn everything he/she can about the agreement negotiated by the field test administrative staff with the local administrative staff – and may propose to re-open those negotiations if he/she believes that his/her organization was not sufficiently involved at the outset.

Specialized Staff

Depending upon the composition of the local school or school district staff, individual members may have intense interest in one or more aspects of the field test. Here are some examples:

Curriculum Specialists. The nature of the product may be of great interest to local curriculum experts. How the product fits with the existing school curriculum will be a concern, for example. The subject content of the new product may overlap, supplement, or displace existing content; the teaching procedures may converge, conflict with, or extend what is already being done. Thus the curriculum personnel will need to be familiarized with the rationale of the product as well as examining its character, but only to the extent that the ultimate distributor (such as a publisher's salesman) would do it.

Similarly, they may be interested in the pupil tests being used because they may contain new ideas for local teachers to adopt in writing tests.

> **CDC**
> Curriculum leaders in the LEAs were particularly interested in the fact that the CDC units contained lesson-embedded test items specific to the performance objectives in those units whereas the special unit tests developed for the field trial were addressed to the more general unit goals rather than to the specific performance objectives. The PSE staff found it necessary to justify testing the goals rather than the objectives.

Staff Trainers. Some local school districts have personnel who specialize in designing and supplying in-service training. Such personnel would be particularly interested in the training techniques to be used in acquainting the teachers with the field test procedures — especially because they may be expected to give such training.

> **CDC**
> Because the PSE staff did not standardize the procedures to be used in training teachers to give the CDC unit pretests and posttests, local training varied widely. Inadequate training in one LEA caused teachers to probe repeatedly for individual children to answer questions on the oral pretest simply because the teachers had not been taught that the pupils were not expected to know the answers on the pretest.

Evaluation Specialists. If there are local evaluation specialists, they will be deeply interested in the field test sampling design and sampling procedures. They may want to know as much about these as the evaluation specialists at field test headquarters and may well want to debate the choice of design and procedures with headquarters personnel.

> **CDC**
> The evaluation staff in one LEA challenged the test administration procedures called for in the field trial plan, saying they doubted that local teachers could use it properly. To check that possibility, the PSE staff went to that LEA, pilot tested the procedures with local teachers, and held a debriefing session with the teachers so that the local evaluators could hear local personnel give their views about the procedures.

The PSE staff learned enough from this to modify the procedures somewhat. In retrospect – looking back over the difficulties many classroom teachers had with those procedures during the field trial– the PSE staffs should have learned even more from the pilot test in that first LEA. The procedures should have been revised to fit the requirements of those teachers who performed least well in administering the instrument during the pilot test rather than to meet the requirements of the typical teacher in the pilot test.

Product-Using Staff

Those who are actually to use the product being field tested will also need to be trained. If the product is an administrative device, the local administrators discussed earlier will also need to be trained. If the product is an administrative device, the local administrators discussed earlier will actually by the product users. In that case, they will need product training as well as orientation to field test administration. But if the product is an instructional device, as is often the case, the users will be classroom teachers, supported as necessary by paraprofessionals, librarians, guidance counselors, and other personnel needed to assist with product use. Product users need to be trained in the following respects.

Nature of the Product. Complete familiarity with the product itself is essential. Its objectives, the equipment and materials needed to use it, the classroom and out-of-classroom procedures called for, the professional and nonprofessional support services called for, the participation by non-school personnel needed, required interrelations of product use with other aspects of the school program – all these must be well understood by those who must use the product.

How to Use the Product. It is not enough that the users understand the product and know why it is being tested. They must also be taught the operational skills needed to use it. This can be best done by demonstrating product use in actual or simulated situations or, better still, by having the trainees themselves practice product use in actual or simu-

lated situations.

Purposes of the Field Test. The specific purposes of the field test need to be explained to the product user. The value of the results to potential product users – of whom those drawn for the field test are a representative sample – should be emphasized. That is, what the field test findings can mean to their counterparts in other school districts, as well as locally in the future, should be described in order to motivate the product users to do a careful job during the field test.

Data Collection. The materials and procedures needed to use the product will be accompanied by a set of materials and procedures needed to collect field data. These must also be taught to the product users, who also must be persuaded that faithful attention to data collection is just as important as faithful use of the product. Indeed, the entire purpose of the field test is to gather such data – not simply to help local school personnel use new educational products. The training should emphasize the fact that the field test instruments must be completed accurately and on time if the field test is to run on schedule.

Report Writing and Distribution. Product users are likely to be interested in their own performance and/or the performance of others (such as students) who engage in the use of the product. They should be taught how to read and interpret the forthcoming reports so that they will not misunderstand them. Given the frequency of "no significant differences" findings in educational research, it is especially important to help product users to interpret the field test findings before they are transmitted to them. Otherwise they may feel that they have somehow failed simply because the product itself is not satisfactory. This is a matter of explaining what is being evaluated: not the skills of the product user, but the performance of the product itself.

Community Relations

The introduction of a new educational product into a local school district has the potential for arousing community interest. The field tester should take a direct interest in whether and how the community is informed because the way it is done can help determine whether the field test accomplishes its purposes.

Some products require that the community be more than informed: they require active participation of local residents as a part of the instructional procedures. For example, many of the new programs for disadvantaged students introduced in the past decade require that the parents of those children assist in their instruction. And the wave of career education programs typically requires that workers visit the school and/or that students visit places of work and interview workers. These requirements must be met if the product is to be used as intended. Initial negotiations with the local site should include careful discussion of the kinds of community participation needed to use the product properly. The capacity, as well as the willingness of the locality to engage parents in instruction, to take students into the community, to bring local citizens into the classrooms, and to modify the school schedule as necessary need to be established before the site is finally chosen.

Apart from the question of actual participation by local residents in the instructional procedures, there is the matter of whether and how the community should be informed about the product being tested. Preferences may vary. The product developers may wish to gain as much attention as possible for the product as a means of furthering its eventual wide-spread adoption. Or, conversely, they may wish to keep the field test quiet until they see the results. Local school administrators may feel that the best way to introduce the product is quietly, with minimum fanfare, so as not to arouse possible public hostility. Or, conversely, they may wish to gain maximum credit as innovative leaders by telling the public all about the test.

The preference of the field tester should be dictated by the general principle discussed earlier: The product should be tested in conditions like those under which it will eventually be distributed. This means that the methods used to inform the community should approximate those that will eventually be employed by the users. Probably the best way for the field tester to produce this condition is to allow the field test sites to do whatever they are inclined to do, neither inhibiting nor stimulating local public information activities. The assumption underlying this recommendation is that since the field test sites are a representative cross-section of the prospective users, they will display a representative cross-section of public relations behavior. These activities may, and of course should, affect product performance during the field test just as they will during later wide-spread use of the product.

What does this general principle imply about the preparation of news releases, slide/tape presentations, handouts for parents, or other devices for informing the community or soliciting participation? It implies that the product developer should prepare and distribute exactly what he/she intends to use during national distribution of the product – no more and no less. And it clearly implies that the field tester, inasmuch as he/she will not be present during national distribution, should do nothing whatever during the field test.

CDC

Inasmuch as no clear plan had been developed for the national dissemination of the CDC units at the time the field trial was run, RDO did not call upon the LEAs to use any particular pattern of informing their communities about the units being tried out. The same is true for community participation: that is, the units did not require active participation for their success. The PSE staff used no initiative, leaving it to the LEAs to operate as they saw fit. The LEAs chose to run the field trial very quietly, partly because they had already engaged in public relations activities on behalf of the entire CDC program and partly because the units arrived late in the school year and the project staffs had their hands full arranging for classroom use.

Monitoring

Monitoring the execution of the field test is essential so that errors can be anticipated and prevented whenever possible or corrected after they are made if necessary. Thus the primary function of monitoring is to prevent mistakes or to correct them promptly. A secondary function is to gather information about field test conditions which is not being gathered through formal instrumentation. Such information can be quite helpful in interpreting the results. An example is the attitude of administrators and teachers toward the field test itself. Their feelings can be decisive in determining whether the product is carefully used and whether evaluation data are conscientiously gathered. Sometimes agreements reached with school district central office personnel in negotiating the field test – and, just as important, the spirit in which those agreements are reached – are not communicated by central office personnel to lower-echelon administrators and to classroom teachers. A third purpose for monitoring is to gather ideas for improving the field testing process itself in subsequent use. Field testing is one activity in which there is no substitute for experience. Monitoring enriches that experience and thus helps the field tester learn more.

Activities to be Monitored

The product being tested as well as the design of the field test dictate the activities to be monitored. However, the following local activities are typical of many field tests.

Sample Selection. The field test management plan may require the local school district staff to select a sample according to instructions issued by field test headquarters. If so, close monitoring of this activity is essential. As mentioned elsewhere in this manual, it is extremely difficult to get local school districts to use random sampling methods

unless there are close supervisory checkpoints. The best monitoring techniques include 1) reviewing sampling instructions with local personnel before they draw the sample and 2) spot checking their work by taking a few instances at random and performing the work independently.

Receipt of Product. Complex instructional systems are increasingly common as instructional products. Delivering the components of the systems to local school districts can be a complex logistical process. Monitoring the actual receipt of all necessary components, which includes careful inspection, logging, and orderly storage, is necessary. This must be done early enough for re-ordering or tracing lost shipments.

Receipt of Evaluation Materials. Field test instructions and instruments may be shipped to local districts from a separate central shipping point and at a separate time from the product to be field tested. Monitoring the receipt, inspection, logging, and orderly storage of these materials is of course just as important as monitoring the receipt of the product itself. (In the case of the product as well as the evaluation packets, the monitor must see that local personnel open all boxes and inspect all packages. Empty envelopes, short counts, duplicate materials, and collating errors must be detected before materials are distributed to classrooms.)

Distribution of Materials. The logistics of getting multiple sets of materials with component parts numbering in the hundreds or thousands from local school district central offices distributed to possibly dozens of buildings and hundreds of teachers are complex. Some local districts will choose to bring all teachers to a central point, instruct them and answer questions, and distribute materials on the spot. Others will use the

school mails or specially-arranged delivery services. Whatever the method, careful monitoring is in order. (Distribution is something that local school districts are accustomed to and it will probably go fairly smoothly, if all materials have been received, checked, logged, and properly stored. Even so, monitoring to see that it occurs on schedule is necessary.)

Actual Classroom Instruction. The actual execution of the field test in the classrooms or other settings for which the product is designed is of course the key factor in determining whether the earlier steps were worth undertaking at all. To monitor the execution requires, of course, that the activity be sampled from place to place and time to time, preferably at random. Monitors should examine the following areas of performance:

1. Is the product being used as intended or with sensible adaptation to local circumstances?
2. Are the necessary facilities, equipment, and materials in place and in use?
3. Are the evaluation instruments being applied as intended and used conscientiously? (This is one case in which adaptation to local circumstances will probably be less welcome than a straight-forward following of instructions from evaluation headquarters.)

Collection of Materials. Unless the full packet of evaluation materials is collected according to plan, key data needed for the final analysis may be missing. For example, the time taken to adminster pupil tests may be a total waste unless teachers also return the class roster requesting demographic information to be used in analyzing the test scores. That is, it is often necessary that the data set be complete for any of the data to

be used. If the management plan calls for forms to be collected serially rather than simultaneously at the end of the field test, the monitor should check this process. Papers allowed to accumulate in the test site may be lost.

Return of Materials. Monitors should check to see that materials are being logged, properly identified, packed, and shipped to evaluation headquarters on the schedule planned. Delay in sending materials back for staggered processing can easily jam the data-processing system at headquarters later on, causing an inevitable delay in completing the field test reports.

Monitoring Techniques

Monitoring may be conducted by any one of the following techniques. They are listed in order of preference and – not surprisingly – in order of expense.

One or more of these monitoring methods may be used from time to time, depending upon the phase through which the field test is moving. For example, on-site monitoring is most important during the execution of the field test; written reports and oral reports can be used with less risk during the preparatory and follow-up stages.

On-Site Staff. The best arrangement is to place full-time or part-time monitors on site. Because they are responsible to evaluation headquarters rather than to local school officials, the monitors can maintain their independence and give honest reports of what is happening in the field without risking their status. They must, of course, build good human relations with the local staff and do their work in an unobtrusive manner so as not to interfere with the field test.

Site Visits. Field visits by evaluation headquarters personnel are highly desirable, no matter what other monitoring techniques are used. Because these visits are relatively rare, other monitoring techniques must be used as well. But there is no substitute for direct, on-site inspection by administrative personnel from headquarters.

Written Reports from the Field. Monitoring can be accomplished by requiring local school district personnel to report their actions at periodic intervals on forms supplied by the headquarters staff. There is an inevitable delay in the submission and analysis of reports and sometimes the event is past before the field test headquarters has a chance to take corrective action. This method of monitoring is, even at best, a poor way to anticipate problems and serves only to record errors. It is of course totally dependent upon the candor of local personnel for its accuracy and upon their conscientiousness for on-time reporting. And it is of course considerably less expensive than having an on-site staff or making site visits. (Site visits by the headquarters staff are especially necessary if monitoring is performed by written reporting.)

Oral Reporting from the Field. Telephone calls to or from the local site can be used as a substitute for or supplement to written reports. Telephone calls have the enormous advantage of earlier reporting but can be considerably more expensive unless there are carefully-built interview schedules that direct the conversation into pre-set channels. Telephone calls are of course the best way of getting emergency information when immediate action is required.

III. DESIGN

The design of the field test grows naturally out of the purposes established for conducting it. As indicated earlier, those purposes are directly related to the information needs of various audiences which must have the answers to certain questions in order to make decisions about the product under test. The field test must be designed to give the most valid, reliable, credible answers to those questions that can be obtained in the time allowed and with the funds available.

Designing the field test requires decisions about site selection, data to be collected, instrument adoption or modification or development, sampling procedures, data-collection techniques, data processing, and data analysis. These seven topics are discussed in detail in this chapter.

Site Selection

There are a number of considerations in choosing sites for a field test. They include getting a satisfactory description of the target locations and populations for the product, determining the number of sites, choosing a sampling method, and determining sample descriptives. The term' "universe" will be used to describe the set of all applicable school districts, while the term "population" will be used when discussing personnel, e.g., administrators and teachers, or students in those districts. Sampling may take place in either or both of these sets. The product itself and the purposes for which the field test is being conducted will detemine where sampling may be used. Later in this chapter a detailed discussion of sampling from the population will be presented. This section deals primarily with sampling from the universe of sites. Because this manual deals with field testing rather than with field trials, we limit the discussion to the case in which evaluators have chosen to use probability sampling – either random sampling or stratified random sampling – in selecting sites and individuals within those sites when sampling is used.

Defining the Universe and Population

Evaluators need to have the product developers identify the universe of locations and populations for which the product is intended. He/she may want to suggest to the developers that they broaden their definition of the target populations to include adjacent groups for which the product may work as well or better than for its primary targets. For example, evaluators may suggest that a product designed for grade 5 be field tested also in grade 4 and in grade 6 to determine whether the developers have graded the product accurately.

Once the developers have defined the universe of intended locations and populations, evaluators should ask them to identify the sub-populations of special interest. For exam-

ple, if the developers are concerned about whether the product will work as well with boys as with girls, as well with bright children as with dull children, with minority children as well as with majority children, they should so state. Evaluators can suggest the variables which may influence learning, based on their own previous training and experience. The variables supplied or selected by the developers can be used subsequently by the evaluators either to construct the sample or to sub-divide it during data analysis.

The universe may be further reduced by eliminating those sites where applicable local, state, or federal laws prohibit the use of the product. But it is actually easier and far less expensive to make such reductions after the sample of sites from the universe has been drawn rather than gathering the data necessary to narrow the universe before the sample is taken. Moreover, there are many factors, not determinable in advance, which will turn out to narrow the universe after the sample has been chosen. Thus the best approach is to oversample suficiently to allow for attrition from any source.

Once the universe has been defined, sampling can begin.

Number of Sites

Two opposing factors are at work in deciding the number of sites: 1)field test costs are lowered and field test management is simplified by keeping the number of sites low, whereas 2) field test reliability, validity and credibility are increased by keeping the number of sites high. For maximum economy and convenience, the ideal number of sites is one. For maximum reliability, validity and credibility, the ideal number of sites is one per classroom required for the field test. But there are other considerations as well that can tilt the balance either way. For example, if the field test is to determine the effect of of the product when all eligible teachers in a given school or school system use it simul-

taneously (perhaps for the psychological effect of universal simultaneous adoption or simply for economies of scale) then universal participation by a small number of sites will supply as much data as the evaluator can collect and process. On the other hand, if the product requires the use of highly specialized equipment or highly specialized personnel, as in the case of a relatively rare high school course in refrigeration, then evaluators will need to select a large number of sites simply to meet the sample size requirements.

Again, if the population of users is to be sub-divided by site characteristics, evaluators need to draw enough sites to keep cell size at an acceptable minimum during data anaysis. In short, determining the number of sites is the result of the complex interaction of a number of factors, some of them conflicting with one another.

Random Sampling

The advantages of random sampling from within the previously-defined universe are well known to evaluators. The method is ideal from the viewpoint of representativeness; it should yield a good cross-section of the kinds of sites for which the product is ultimately intended. This allows for virtually unrestricted extrapolation of the results to the full universe, with qualification.

On the other hand, as is also well known to evaluators, straight random sampling can produce a set of sites which are inconveniently located, uncooperative, incompetent to carry their share of work in executing the field test, and in other respects difficult for field testing. These are, of course, exactly the difficulties with which probability theory wishes to saddle the evaluator. If he/she wants to extrapolate to the universe, then he/she must deal with an authentic sample of the universe. Given the high likelihood of severe attri-

tion from a number of sites chosen at random, oversampling should be ample to compensate for the loss of locations.

Another difficulty with random sampling is that the cells intended for use in the analysis may not contain a sufficient number of cases. This can be prevented in advance by using stratified random sampling, employing the variables of interest selected by the developers or evaluators, either because they are thought to influence learning or because they measure key target populations about whom the developers are especially eager to have performance information. These variables can be used to stratify the universe of sites before the sampling is taken, assuring sufficient cell size for the intended data analysis. This is yet another illustration, among dozens that can be cited, of how field test activities interlock with each other. In this case, the purpose of the field test conditions the method of analysis which conditions the desired sample characteristics which conditions the site selection process.

Sample Descriptors

Site characteristics thought to influence learning can be used either to stratify the sample in advance, as mentioned above, or to describe it after it has been drawn. In either case, data on such variables can be invaluable in explaining product performance and should be collected during the field test.

An illustrative listing of such variables is included because they may be used to stratify the sample. The same list can be used in the following section of the manual, to illustrate descriptors that can be gathered to explain product performance afterward if not to stratify the sample in advance.

Community. Geographic location, urban/suburban/rural, size, wealth, degree of industrialization, growth rate, etc.

Parents. Income level, years of education, racial and cultural characteristics, occupations, etc.

School System. Number of schools, per pupil expenditure, history of innovativeness, previous field test experience, quality of leadership, etc.

Schools. Physical facilities, number of professionals and paraprofessionals per 1,000 pupils, professional specialists, etc.

Curriculum. Relation of product content to existing curriculum, relation of product instructional techniques to current instructional techniques, etc.

Staff. Training, years of experience, racial and cultural characteristics, etc.

Students. Ethnic composition, level of ability, socio-economic status, etc.

CDC

The primary reason for designating the field assessment of CDC units a "field trial" rather than a "field test" is the fact that the ten LEAs serving as test sites were not a random sample of the universe of locations for which the CDC units were intended.

In their capacity as product developers, RDO and the LEAs indicated that the CDC units were designed for use in all of the approximately 17,000 school districts in the nation. While the ten LEAs participating in the project were somewhat representative of the 17,000 in many respects (see list of sample descriptors immediately above) the ten were decidedly not representative in the intensity of interest and involvement in a structured cognitive curriculum. Indeed, the ten were selected from among the 17,000 precisely because of their outstanding records in structured cognitive curricula.

The ten were chosen as field trial sites primarily for reasons of economy and administrative convenience. While the field trial findings cannot safely be extrapolated from these sites to the 17,000 (one purpose of running the field trial) they were highly suitable for getting leads to unit revision (the second purpose of the field trial). Indeed, their maturity and extended experience with cognitive curricula made them particularly good choices for the second purpose. Their project staffs, and their classroom teachers, could be counted on to be cooperative and responsive in making critiques of the CDC units and offering suggestions for improving them.

The fact that ten sites were to be used for the field trial was an automatic extension of the fact that ten sites had created the CDC units. And even that number yielded to other pressures part way through the field trial. When the funding agency ordered the field trial design changed from 60 pretest – posttest classrooms without

control groups to a pretest – posttest design with 40 experimental classrooms in mid-March, of the field trial year (near the end of the school year with 27 units yet to enter the classrooms), the LEAs demanded some offsetting reduction in their responsibilities. This was achieved by reducing the number of CDC units tested at each site from ten down to five. That is, the LEAs had to supply 33 percent more classrooms but tested 50 percent fewer units.

Despite the fact that the sites were not chosen at random, elaborate descriptions of site characteristics, staff characteristics, and student characteristics were gathered, analyzed, and used to interpret the performance of the CDC units.

Sampling

This manual does not cover in detail the complex subject of sampling. Rather, some general observations are made on the sampling issues an evaluator must address as he/she plans the field test.

Since field testing almost always involves sampling – usually it means drawing only a tiny sample from the complete universe or population – it is essential that field testers have a good grasp of the subject so that they use the most efficient sampling procedures and get maximum information from the relatively small samples they are typically forced to use.

Complete Census Versus Sampling

A field test can be designed in the form of a complete census of the population of interest only if the universe is quite small, if sufficient time is available to reach all members, and if cost is no problem. These conditions almost never exist, but they are conceivable.

For example, a product intended to teach beginning reading skills to Indian children of a small tribe that speaks a unique dialect might be field tested with all of the several hundred six year old children of the tribe if they are conveniently enrolled in a limited number of accessible schools. (From a statisical point of view, the existing six year old children are actually a sample of successive waves of six year olds yet to come. It is the effectiveness of the product for those future six year olds to which the field test is addressed; therefore, tests of statistical significance are appropriate to extrapolate the findings to these future children.)

In almost every case, however, field testers will find it necessary to sample the currently-available members of the universe and to extrapolate the results to other current and future members of that population.

> **CDC**
> Even in the ten LEAs which had produced the CDC units, there were extremely strict limitations on sample size. The ten LEAs were scarcely able to supply the minimum of 60 classrooms for the early units. Many teachers drawn at random for the field trial had to be personally persuaded to use the units. When the quota was raised from 60 experimental classrooms to 80 classrooms (40 experimental and 40 control) the LEAs reacted by insisting on field trials for only half the units originally assigned to them.
>
> There were, of course, reasons for this resistance — late delivery of units to the schools, the burden of using the heavy instrument package and so on. But there are always reasons for such resistance in any field trial, especially one using random selection. Under such conditions there are strict practical limits on sample size. Moreover, even with the high evaluation budget of the CDC project, the cost of a complete census would have been far beyond reach.

Forms of Sampling

There are many different forms of sampling that field testers can use, once they have defined the universe and population. As explained briefly in an earlier discussion, random sampling is highly advantageous in that it yields a fully representative cross-section of the population for which the product is intended and thus allows for extrapolation to the full

universe. But, as also explained in that discussion, random sampling can produce an extremely inconvenient set of subjects: some of them geographically scattered, uninterested, uncooperative, incompetent, and so on.

To avoid this problem, field testers can reduce the universe in a number of ways before taking the sample. For example, they can circulate a memorandum to prospective sites describing the planned field test, stating what will be required of the sites selected, and soliciting applications from interested locations. Any such procedure, of course, introduces an immediate bias into the sample of sites. In this example, the universe will be reduced to that set of school districts which are interested in the product, willing to meet the conditions imposed by the field testers, and energetic enough to apply. The results of the test can then only be extrapolated to such school districts. Field testers will have to consider the gains in convenience and economy they can accomplish by reducing the universe through such techniques versus the losses in their power to extrapolate their findings to the full universe. Field testing is always an exercise in extrapolation to some population not included in the test: reducing the power to extrapolate always reduces the utility of the field test results to some degree.

Another difficulty with random sampling, also mentioned earlier, is that the sample drawn may contain an insufficient number of cases representing a particular variable of interest. For example, if the effectiveness of a product for a minority subpopulation which is having difficulty in school is a question of particular interest in the field test but the size of that minority subpopulation is quite small in a given school district or state, random sampling the entire universe may not turn up enough members of that particular subpopulation to allow a judgment about the effectiveness of the product for them. In such cases, stratified random sampling can be used. That is, the universe can be sub-

divided before the sample is drawn and enough cases taken from each subdivision (one such subdivision might be the members of the minority subpopulation mentioned above) to allow conclusions about the effectiveness of the product for the population in each subdivision.

Unless some form of random sampling is employed, the bias introduced into the sample will be indeterminate and it will not be safe to extrapolate to any population since the universe from which the biased sample was drawn will not be known.

Ideally, having drawn the sample, field testers will be able to make random assignments of subjects to the treatment. That is, they will be able to decide at random which students or classes or schools or school districts will use the product and which will not. This method is necessary for any kind of design involving a comparison of experimental and control groups in which it is essential to establish comparability between the two groups at the beginning in order to draw valid conclusions about the differences between them at the end. That is, if the effectiveness of the product is to be judged by comparing the achievement of the group using the product to the achievement of another group not using the product, the comparability of the two groups must be established. Dividing the total sample at random into experimental and control groups is the best way of doing this.

There are a number of statistical methods for "equating" two groups which have not been chosen at random from a larger group, but the equating must always be done using a finite number of independent variables. Since the variables chosen may or may not include all the independent variables that influence product outcomes, the field testers cannot be certain that the groups have in fact been equated. Thus statistical procedures

for equating two groups cannot be used in complete safety.

Local school districts can be expected to resist the idea of random assignment to treatments. They usually reason that if a product is worth using (or trying) with some students, it is worth using (or trying) with all. Moreover, they tend to introduce innovations by drumming up as much enthusiasm as possible among classroom teachers, a technique which does not square with allowing only some teachers to use the promising new product. Field testers should be prepared to encounter this problem and should consider making random assignment to treatment a condition of participating in the field test. Such a condition, like any other condition set by field testers, tends to bias the sample by limiting it to school districts that will accept such a requirement. Those districts, of course, may not be fully representative of the entire universe of school districts, a factor that must be taken into account when the field test results are interpreted and reported.

CDC

All ten LEAs resisted the idea of drawing teachers at random either to use the CDC units or to serve as controls, i.e., administering tests without using the units. Despite the fact that these LEAs had been singled out as national leaders in career education and were as positive toward the program as any district might be to anything new – indeed, they had to a large extent constructed the program – they resisted random selection. They expressed great reluctance about announcing to individual teachers that they had been chosen to teach the units. They explained that such a procedure could set back the cause of a structural cognitive curriculum in their schools, a cause that they had been working for years to advance. Although the problem was most severe in the LEAs with strong teachers' unions where local administrators walked with greater care, it existed in each one of the ten sites.

Size of Sample

Sample size should be determined in part by the size of the total population. The larger the population, the smaller the sample size can be as a proportion of the total – and the

smaller it must be, in the interests of economy. Conversely, the sample can constitute a larger proportion of smaller populations – and may have to do so for statistical reasons and credibility.

The sample must be large enough so that tests of statistical significance will not fail to show differences simply because the sample is too small to yield results at the selected levels of significance. On the other hand, it is wasteful to make the sample so large that even trivial differences in scores show up as being statistically significant simply by virtue of the huge number of cases available to verify those differences. For example, if the average difference between the group using the product and the group not using it is ten points in a valid and reliable 30 -point test, the sample should not be so small that such a difference is declared non-significant. But if the average difference between the two groups is ¼ point on the same test, the sample should not be so large that the difference is declared significant.

Apart from the question of statistical significance, the sample must be large enough to lend credibility to the findings. That is, it must seem "reasonable" in size to the professionals who must make decisions based on the field test results. Generally, results for several hundred individuals for 25 or more classes or for ten or more schools or for three or more school districts are at the minimum threshold of credibility. That is, smaller numbers, even if they show statistically significant differences, can be shrugged off as being non-representative and as having produced results that might be different had the product been used with other people.

CDC

The five to ten districts and the 40 to 60 classes using CDC units represented minimum figures for summative evaluation and extra-

polation to other sites. Indeed, the credibility of the figures may be below minimum to prospective adopters of the CDC units.

On the other hand, as explained earlier, the samples were needlessly large for formative evaluation and the eliciting of ideas for unit revision.

Unit of Sampling

As indicated above in a discussion of stratified random sampling, the sample must be large enough to yield a sufficient number of cases for conclusions about each subpopulation of special interest. That is, when subdivided into cells to analyze the results for one group against another (boys versus girls, whites versus blacks, bright students versus dull students) there must be enough students of each type to support conclusions about each group and to support comparisons between them.

When a universe is stratified for the pupose of drawing the sample, but those stratifications are not retained in the analysis (for example, when urban, suburban, and rural school districts are included in the sample but the results are to be analyzed for all districts combined), field testers must choose between taking the same absolute number of cases from each subdivision (for example, urban, suburban, rural) or taking numbers proportional to the size of those subdivisions in the total universe.

For example, if 3,000 cases are to be drawn, shall 1,000 be taken from urban school districts, 1,000 from suburban, and 1,000 from rural districts or shall 1,200 be taken from urban school districts, 1,500 from suburban, and 300 from rural districts because those are their proportions in the total universe? The choice depends on whether the field testers want to make certain they can speak with safety about each segment of the total universe (even though results for those segments are not analyzed separately) or whether they want to speak about the total universe, even though the findings may not fit every segment of the total.

One critical aspect of sample size is the definition of the unit of instruction. This problem has been discussed in the professional literature previously.

If a product is used in a classroom of 30 children, do the characteristics and the skills of the teacher interact with the product in such as way that the class constitutes only a single case rather than 30 cases, or does each student have an independent experience with the product so that there are 30 cases for statistical analysis rather than one case? Field testers must make a conscious resolution of this issue in selecting a sample size. Otherwise they may find that critics of their work will reduce their sample size to 1/30 of what they intended it to be or expand it to 30 times what they thought they had gathered. The first interpretation could rob the findings of all statistical significance while the second could make their sampling design look unnecessarily expensive because of the surplus cases.

The choice turns upon the character of the product being field tested. In general, products designed for mass use in a classroom tend to cause the class to operate as an indistinguishable unit, with the characteristics and skills of the teacher interacting with the product to produce learning and with the actions of each individual student having some influence on the learning of all other students. In such circumstances, the classroom can be understood to constitute one instructional unit. In contrast, if a product is designed for highly individualized use by the student, if the teacher plays a minor or non-existent role in instruction, and if the actions of one student do not influence the learning of another, each student using the product can be regarded as one unit of instruction.

CDC

Inasmuch as the CDC units were designed for mass use in classrooms,

the entire classroom was regarded as one unit of instruction and the sample size set at 60 or 40 experimental and 40 control classrooms accordingly.

Attrition

Attrition must be expected during field testing. Equipment and materials will not arrive in time; administrators will lose interest; teachers will become ill; classes will be broken up in mid-year and be reassigned to other teachers; the year will end before the teacher gets to the final lessons in the product; respondents will supply some data but not others, producing incomplete data sets that cannot be fully analyzed; results from some schools and classes will arrive late, delaying the analysis of the results from other schools and classes; and so on. The solution to all these problems is over-sampling -- taking a sample larger than is actually needed to supply the amount of data required for the analysis of results. Loss of data from attrition can be assumed to range from 10% to 20% of the total, for a variety of reasons.

While over-sampling can assure the field testers of getting enough data to analyze, it does nothing to handle the problem of bias introduced into the sample by attrition. Because it cannot be assumed that attrition occurs at random -- that is, those who fail to supply the necessary data differ in characteristics and skills from those who do supply the necessary data -- the field testers must attempt to identify and describe the kinds of bias introduced by attrition and use that information when interpreting the findings. For example, if it should appear from the reports of the field monitors that the teachers who did not finish using the product as intended were less energetic and less competent than those who did, the field test findings will apply only to classes taught by unusually energetic and competent teachers. This must of course be reported for the guidance of those who sponsored the product, those who developed it, those who must revise it, and those

who may wish to use it.

CDC
An attrition rate of 16⅔ percent was assumed for the CDC field trial. That is, it was assumed that a sample of 60 field trial classes would yield 50 sets of useable data and that a sample of 40 classes would yield 33 sets of useable data. This attrition allowance proved adequate for 22 of the 25 units, with the remaining three falling just below quota.

Drawing the Sample

Field testers must define those who are eligible to participate in the field test in each cooperating local school district. The districts themselves tend to define "eligible" as equivalent to "interested". That is, they tend to assume that teachers who are not interested in using the product would be automatically eliminated from any list of the population to be sampled. Such an elimination of course biases the sample immediately and in unknown ways, endangering the ability of the field testers to extrapolate the results to the universe.

The assumption that only the interested should be considered eligible for participation is so automatic in local school districts (and quite natural for them, since it fits their typical approach to introducing innovations) that the field testers themselves must write a careful definition of eligibility and monitor local personnel to see that the official definition is used.

Field testers should anticipate that school districts will propose that various classes of teachers are ineligible and should not be sampled for the field test. These may include substitute teachers, teachers who have severe personal problems outside the school, teachers already participating in other innovative programs, teachers not expected to

finish the school year for one reason or another, teachers who resist administrative leadership in any form, teachers who reject all innovations, and so on.

The field testers must decide whether they wish to eliminate such classes of teachers from the eligibility roster before drawing the sample or whether they wish to include them in the roster, take the chance that their names will be drawn, and live with the consequences during the field test.

Identical considerations of course apply to school building administrators, who may for one reason or another declare their schools to be ineligible for the product and thus eliminate themselves from the sample. Any categories of teachers or administrators (or pupils, such as the mentally handicapped) eliminated from the eligibility lists automatically reduces the universe to which the eventual findings can be extrapolated. Thus determining eligibility is a matter of serious concern for field testers and must not be left to the discretion of local school officials.

CDC

The PSE staff issued the following instructions to the LEAs:

1. Teachers may be removed from the eligible list only for reasons of:
   a. long-term illness
   b. leaves of absence which will interrupt the teaching of a unit
   c. separation from employment which will interrupt the teaching of the unit

2. Teachers cannot be removed from the eligible list for such reasons as:
   a. confidence
   b. interest in the new program

In short, the only reason regarded as acceptable for declaring a teacher ineligible to be drawn for a CDC field trial unit was his/her absence from the classroom during the period the unit would be taught.

If a product comes in segments to be taught separately so that the same teacher need not teach all segments, field testers must decide whether to use sampling with replacement or sampling without replacement and must write the sampling instructions accordingly. That is, field testers must decide whether a teacher who has used one segment of a product is to be automatically eliminated from the sampling pool when names are drawn to teach other segments of the product. Local school officials may feel that a teacher has "done his/her share" when he/she has taught one segment and should not be asked to teach another until every other teacher has had his/her turn. Following that procedure means using sampling without replacement. Field testers need to remember that statistical formulas differ according to which procedure is used in constructing the sample. If the samples are drawn locally, they must write specific instructions for all local school officials in order to get consistency across all sites so that the data can be pooled.

Samples can be selected by the central field testing staff, using rosters of eligible students, classes, teachers, or schools supplied by local school district administrators depending upon the definition of the unit of sampling. This is the best procedure for ensuring uniformity across all sites and adherence to the sampling plan. However, the method entails one psychological difficulty and one procedural difficulty. The first is that local school officials may feel that an outside authority is issuing orders to local teachers (even by drawing their names at random in a publicly-described procedure) and thus may be in violation of local control of education, local administrative prerogatives, or local teachers' union contract. The second problem is primarily logistical. It is obviously time-consuming for local school districts to supply complete rosters of all students, classes, teachers, or schools in convenient form for easy use by headquarters staff.

Samples can of course be drawn locally under procedures written by the central headquarters staff, which gets around the psychological and logistical problems mentioned above. Of course, local sampling creates its own problems – and they can be worse. Even when procedures are very precisely written, which they of course must be, local school district officials cannot be relied upon to carry them out faithfully, even if they understand the instructions perfectly. They will have compelling reasons to eliminate some students or teachers or schools from the sample, either before or after it is drawn, in order to simplify local administrative problems and minimize disruption caused by the field test. Active on-site monitoring of local administrators who are doing the sampling is absolutely essential to make sure that instructions are followed – or more realistically, to record deviations so that they can be accounted for when interpreting and recording the final results. No matter how well written the sampling instructions are, oral review by telephone, or preferably in face-to-face meetings with those who will draw the samples, is essential. Ideally, on-site monitors will first teach the procedures to local officials and then run spot checks to see that they are faithfully executed.

Sampling procedures should of course be covered during initial negotiations regarding the participation of the site. But even this will not assure that the sampling methods are known to and accepted by all the lower-echelon persons, including school building administrators and classroom teachers, who must ultimately agree to them.

CDC

Despite the fact that the PSE staff issued 11 pages of tightly-drawn instructions to the LEAs describing how they should draw the field trial sample classrooms, the procedures were not always followed. lowed.

During the course of the field trial, the PSE staff found in one LEA that the assistant superintendent of schools had been making it a standard practice to review the names and schools of all randomly-

Samples can of course be drawn locally under procedures written by the central headquarters staff, which gets around the psychological and logistical problems mentioned above. Of course, local sampling creates its own problems – and they can be worse. Even when procedures are very precisely written, which they of course must be, local school district officials cannot be relied upon to carry them out faithfully, even if they understand the instructions perfectly. They will have compelling reasons to eliminate some students or teachers or schools from the sample, either before or after it is drawn, in order to simplify local administrative problems and minimize disruption caused by the field test. Active on-site monitoring of local administrators who are doing the sampling is absolutely essential to make sure that instructions are followed – or more realistically, to record deviations so that they can be accounted for when interpreting and recording the final results. No matter how well written the sampling instructions are, oral review by telephone, or preferably in face-to-face meetings with those who will draw the samples, is essential. Ideally, on-site monitors will first teach the procedures to local officials and then run spot checks to see that they are faithfully executed.

Sampling procedures should of course be covered during initial negotiations regarding the participation of the site. But even this will not assure that the sampling methods are known to and accepted by all the lower-echelon persons, including school building administrators and classroom teachers, who must ultimately agree to them.

CDC

Despite the fact that the PSE staff issued eleven pages of tightly-drawn instructions to the LEAs describing how they should draw the field trial sample classrooms, the procedures were not always followed.

During the course of the field trial, the PSE staff found in one LEA that the assistant superintendent of schools had been making it a standard practice to review the names and schools of all randomly-

drawn field trial teachers and changing the names so as to level out the field trial work load over all teachers and all schools in the district. The effect was identical to what random sampling would have accomplished, given enough samples drawn for enough units. But like most local administrators, this one trusted his own judgment more than the laws of probability.

Data

The data to be collected are those necessary to achieve the purposes of the field test. In general, this means the data needed to answer the questions for which various classes of decision-makers will use the field test information. Evaluators will need to interview the intended audiences for the report in order to know what decisions those audiences must make based in whole or in part on the field test results. Evaluators are likely to find that there are many more questions than there is money to find answers. Thus he/she must organize and display these to his/her funding agency, indicating what persons have asked what questions, showing the costs of answering them, and getting the funding agency to rank the importance of answering specific questions for specific audiences at specific prices.

The evaluator may want to give his/her own views on the value of answering the various questions for the various audiences and offer his/her funding agency guidance in deciding what questions he/she is to answer. Five types of data will probably be required to answer audience questions.

Site Characteristics

The Site Selection section contained a list of site characteristics of the kind that may be relevant to product performance. Data of this kind will be needed for extrapolating the findings to the universe from which the sample was drawn and for associating outcomes with site characteristics.

Conditions of Use

Perhaps the most salient data to be gathered are those that will describe, reliably and in detail, the actual conditions under which the product was employed. Such data are essential to determine whether the product was used as intended by its developers and, beyond that, to find whether differential performance was influenced by the conditions of use.

The categories of information will probably include the following:

1. *Completeness of Use.* Did teachers use all of the product or did they omit certain segments?

2. *Schedule of Use.* Did teachers use the parts of the product in the intended sequence; did they compress the schedule from what would normally be expected; did they extend the schedule beyond the normal expectations?

3. *Augmentation.* Did teachers supplement the product with extraneous materials or inventions of their own design to make it more effective?

4. *Training.* Did teachers receive the advance training called for by the product developers?

5. *Physical Facilities.* Were the spaces in which the product was used and the equipment employed with it what the originators intended?

6. *External Support.* Were the administrative arrangements and the availability of specialized support personnel such as librarians or guidance counselors what the developers imagined they would be?

7. *Access to Supplementary Locations.* Did the teachers manage to make any special field trips or other use of outside resources?

Attitudes Toward the Product

The attitudes of those who use the product during the field test may turn out to be highly predictive of the reception the product will meet if it is scheduled for general distribution following the field test. Therefore, the attitudes of the users should be studied carefully during the field test itself. For an instructional product, there are at least four key populations whose attitudes can be pivotal in deciding the eventual acceptance of the product:

1. *Administrators.* What did the school superintendent, other central office personnel, and school building administrators think of the product: its importance, its utility, its administrative convenience its probable effectiveness, and the likelihood that they would want to maintain it after the field test ends?

2. *Teachers.* What did the teachers think of the product: its importance, its quality, its utility, its fit with the existing school curriculum, its probable effectiveness, and the likelihood that they would want to continue using it after the field test ends?

3. *Students.* What did the students think of the product: its importance, its attractiveness, its appeal relative to their other studies, and its probable effectiveness?

4. *Parents.* What did the parents think of the product: its importance, its motivational value for pupils, its appeal for students relative to their other studies, and its probable effectiveness?

Effect on School Program

Some products do not impinge on the operation of the school program outside the grade level and subject field touched by the product being tested; others may have a tangential relationship to other grades and subjects; while still others may operate to support or perhaps to disrupt other aspects of school operation. For example, a product designed for individualized use may improve the climate for individualized procedures and materials throughout the school, having a favorable effect on the use of similar products in grades and subjects other than those engaged in the field test. On the other hand, a product requiring pupils to leave the school building for more than a single class period may require changes in the school's schedule and may compete with or do actual damage to student learning in the grades and the subjects not engaged in the field test. Thus data should be gathered to show the positive or negative effects of product use on other aspects of school operations.

Pupil Performance

Assuming that the product being field tested is an instructional product intended to produce changes in pupil behavior, data on pupil performance are, of course, fundamental.

The performance areas may include cognitive, affective, or psychomotor, according to the intended effects of the product.

Depending upon the evaluation design, data may need to be collected both before and after the product is used (pretests and posttests) and for groups who did not use the product as well as for those who do use it (control groups versus experimental groups).

Performance scores may be reported in terms of relative growth (before-treatment standing versus after-treatment standing) or in terms of position on a normative standard (before and/or after treatment) or in terms of an arbitrarily-set criterion of satisfactory performance (before and/or after treatment).

Information on student characteristics must be gathered to determine whether the product works equally well for all kinds of students. For any particular product, specific student characteristics may be expected to be learning-related. A clear-cut example would be a product designed for students with severe physical or mental handicaps. Observation or measurement would need to be taken to establish the degree to which students have the specified handicaps and to relate their presence to student learning. More generally, the effectiveness of many products will be associated with differences in students' level of intellectual ability, ethnic background, socio-economic status, level of parental education, and other demographic indicators commonly associated with school achievement.

CDC

Data were gathered on site characteristics, conditions of use (as reported by teachers themselves), teacher and student attitudes toward the units, and pupil performance associated with pupil characteristics. Data were not gathered on the effect of the units on the school program, but probably should have been (although they would have been greatly influenced by teacher annoyance at the late

arrival of the units) and data were not gathered about the attitudes of administrators and parents toward the specific CDC units because their attitudes toward the total CDC program (gathered earlier and found to be quite favorable) were thought to be better indicators of their general behavior than their attitudes toward specific units. The data were sufficient for measuring pupil performance and associating it with pupil characteristics, but were not sufficient (until supplemented by other data gathered by RDO) to guide unit revision. Thus, while serving the summative purpose, they were not sufficient for a formative evaluation purpose.

The PSE staff as external evaluators joined with the RDO internal evaluation staff to develop an elaborate teachers' questionnaire to get information on how the units were actually taught and what the teachers thought of the units. Data from the joint questionnaire were used for both summative and formative purposes. The joint effort represented one of the few completely successful attempts -- out of many made -- to find a common ground between the interests of the external evaluators and those of the internal evaluators.

Instruments

The instruments needed for the field test are those necessary to collect all the data required. Putting together the necessary instruments will be a time consuming and expensive process.

There are many considerations to be taken into account. The primary ones are discussed in this chapter. However, field testing any product will create unique data requirements and perhaps unique instrumentation challenges. Therefore, the following discussions must be regarded as general and illustrative rather than as exhaustive.

Review Purposes of Field Test

It is wise to review the purposes of the field test at this stage. The process of identifying, modifying, creating, and assembling instruments almost always suggests additional kinds of data that might be gathered. So as to avoid gathering data simply because it is possible to do so, the field testers must keep the purposes of the field test firmly in mind and should include in the data set only the information required to accomplish the field test

purposes. Each bit of additional data included in the analysis raises the cost of the work; it of course makes no sense to gather data which the field testers cannot afford to analyze.

Review Types of Data to be Collected

As previously implied, the set of instruments must be sufficient to gather all necessary data on site characteristics, conditions of product use, attitudes of users and non-users toward the product, effect on the school program, pupil performance, and other data needed to describe the circumstances associated with product success. As indicated above, work on the instrument package is very likely to suggest possible expansions in the data set. Such expansions must be considered thoughtfully rather than adopted automatically simply because it is feasible to adopt them.

Re-examine Product to be Tested

A further inspection of the product itself at the instrument development stage can suggest instrument content (data to be collected) or instrument formats or convenient and economical techniques of instrument administration. Among the product features to be inspected are the following:

1. *Product Objectives.* If the product is intended to accomplish general goals stated at a high level of abstraction rather than specific objectives stated at a low level of abstraction, that fact will have clear implications for the kinds of instruments to be constructed. In general, the more specific the objectives of the product the more likely it is that new instruments will have to be developed. Conversely, the more general the objectives the more likely it is that existing instruments can be adopted or modified as measurement devices. This has direct implications for the cost of instrumentation.

2. *Grade Levels.* The grade levels specified for product use tend to determine whether verbal or non-verbal test items will need to be created and whether verbal items are to be in oral or written form.

3. *Subject Fields.* The nature of the subject content being transmitted by instructional products will tend to determine the kinds of abilities to be tested. For example, an instructional product in science will tend to call for skills in data analysis while an instructional product in art will tend to call for the ability to make aesthetic distinctions.

4. *Target Population.* The mental abilities, ethnic composition, and language backgrounds of the populations for which the product is intended will influence the vocabulary and syntax in which verbal items can be expressed (or even the language in which they are to be cast) as well as the assumptions that can safely be made about the pre-existing knowledge and experience the subjects will bring to the field test. One type of experience, the previous test-taking experience of the teachers and students to whom the product is directed, may tend either to limit the kinds of items to be developed or that added time be allowed for them to learn how to handle the unfamiliar items.

5. *Group Size.* Products intended for individual use rather than for use by small groups or large groups may need to be tested in individual or group settings, as appropriate. This has clear implications for the kinds of instruments as well as administrative procedures that can be employed.

6. *Schedule of Use.* The schedule for product use will tend to influence the schedule for test administration. For instance, if the product requires work experience for students outside the school, and if direct observation of student performance is required as a measure of product success, data collection instruments will have to be administered at the work site during the time in which the work experience takes place.

7. *Single Product Versus Member of Set.* A single product which must accomplish its purposes alone may require a different kind of instrumentation from a product to be used as one of a large set where the cumulative effects of the set are of interest rather than the contribution of a single element.

8. *Links to Other Aspects of School Program.* A product to be used in isolation can be evaluated with less comprehensive instrumentation than a product intended to influence or to be reinforced by parallel events elsewhere in the school program.

9. *Previous Evaluation History.* A product that has been through careful inspection or simulated use or pilot trials or pilot tests or field trials will have accumulated a set of instruments that can be quite helpful in yielding design ideas for the field test instruments. In fact, earlier instruments may be entirely suitable for re-use in the field test. Moreover, earlier evaluative findings will suggest product effects that can be searched for during the field test.

10. *Build-In Data Gathering Devices.* The use of certain products automatically yields data without a specific intervention by the field tester. For example, students may produce work samples during product use which can be collected and analyzed as a part of the field test data set. Such data require no special instrumentation to collect. But plans should be made for analysis of the products to be sure that their use will yield viable data.

List Criteria

Field testers may find it useful to distinguish between what may be called *internal* criteria and what may be called *external* criteria for product performance. Internal criteria are defined in this manual as those used in determining whether the product has achieved its own objectives; external criteria are defined in this manual as those used for determining performance compared to the performance of alternative products with identical or very similar objectives.

Internal Criteria. Three kinds of criteria are frequently used to determine whether a product has accomplished its own objectives:

Product Effectiveness. Did the product accomplish its primary objectives? In the case of an instructional product, did it change pupil learning as intended? For what populations and under what conditions of use was this accomplished?

Product Utility. Was the product easy and convenient to use or did it require extraordinary energy outputs by the staff or unusual administrative arrangements to make it successful? Did the product have any positive or negative effects elsewhere in the school program that would make it especially attractive or unattractive?

Product Acceptability. How favorable or unfavorable were the reactions of administrators, teachers, students, and parents to the use of the product and what were their estimates of its effectiveness?

Specifying the criteria to be used in judging product performance is an essential final step before undertaking the selection and development of data-gathering instruments.

External Criteria. Several kinds of criteria can be used to determine whether a product has performed well enough in relation to alternative products to justify further investment and/or general dissemination. However, these criteria cannot be employed unless the field testers have at hand, or can easily acquire, comparable data for parallel products. Given the limited amount of field testing that has taken place thus far, such comparable

data will seldom be available. When it is, the external criteria could include these:

Product Effectiveness. The extent to which the product achieved its intended objectives in improving student learning can be compared to the degree to which other products have been able to produce similar kinds and amounts of learning.

Product Utility. The comparative convenience of use of the product and its effects on aspects of the school program outside the immediate setting of the field test can be compared to those for similar products where the data are available. In general, ease of use and lack of entanglement with other parts of the school program are favorable for the widespread adoption of a product.

Product Acceptability. Attitudes toward the product on the part of administrators, teachers, students, and parents compared to their attitudes toward alternative products can be pivotal in determining whether a product is destined for widespread adoption. Comparative analyses should be made wherever similar data on product acceptability are available.

Effort Required. A product should be above-average in effectiveness if it requires above-average effort to use. The amount of effort required to use the product effectively will tend to show up in measures of product utility and product acceptability but is worth singling out for separate treatment since the effort required to use a product initially is usually greater than the effort required to continue using it. Where comparable data are available, testers should estimate the amount of effort required for continued use of the product in relation to alternative products.

Time Required A product achieving equal effects in less time than an alternative product is likely to find a greater acceptance in the market, so long as its cost and the effort required to use it are not inordinate. Thus comparative analysis of time requirements should be made where comparable data are available.

Cost/Benefit Ratios. The dollar cost of acquiring and using a product in comparison to the student learning or other benefits it produces is of course a matter of critical interest to all parties. Unfortunately cost/benefit ratios are ordinarily not available for alternative products. However, where they can be acquired for even one alternative product, they should be computed for the product under field test, given the importance of such comparisons.

Re-Examine Data Processing Capabilities

Field testers should project their data processing capabilities and use these as a constraint on the kinds of instruments they develop. Although it is always possible to process any kind of data, given enough time and money, it is sensible to limit the data to the types that can be handled within the time and money available. Thus access to test scoring

machines, raters of written compositions, judges of observational data, efficient manual scorers, desk calculators, and computers all should be taken into consideration before instruments are developed.

This is perhaps as good a place as any to remind the reader of what was said much earlier: Field testing is an activity that absolutely requires detailed planning. Every step along the way, from initial site selection to final report distribution, must be thought through in advance in order to schedule operations, estimate staffing needs, and formulate a budget.

Re-Examine Reporting Requirements

Before instruments are selected, modified, created or assembled, field testers should study their reporting requirements and see what these imply for the design of the field test package. The implications can range from substantive content that may have been forgotten in earlier planning to simple conveniences in instrument format. For example, if field test reports for certain audiences are to contain only certain data, it may be wise to collect those particular data with separate instruments for convenient analysis and easy reporting.

Review Literature and Existing Instruments

There are three possible benefits in reviewing the literature and searching through existing instruments before undertaking the development of new instruments:

1. Existing instruments may be usable or may be readily modifiable for use in the field test;

2. Types of instruments previously unknown to the field testers may turn up and, even if they cannot be used, may suggest what kind of instrument can be developed; or

3. Alternative ways of administering instruments may appear in the literature and give useful ideas to the field tester.

Any one of these potential benefits fully justifies the time spent in reviewing the literature. The research abstracts available through ERIC and the standard reference works on measurement instruments will, of course, make the job more efficient.

Select Data Sources

Once the kinds of data have been determined, the populations from whom the data should be gathered can be selected on the basis of accuracy, economy, convenience, cooperativeness, and so on.

Specify Plan for the Instruments

Having carried out the eight steps called for above, the field testers will be ready to specify the plan or table of specifications for the instruments. Those specifications must include the following information:

1. *Types of Data.* An exhaustive list of all data to be gathered either through direct instrumentation or as a natural yield from normal product use.
2. *Types of Instruments.* Observation guides, interview schedules, questionnaires, teacher or pupil diaries, criteria for judging performance situations, paper and pencil tests, work samples, anecdotal record forms, and so on.
3. *Types of Items.* Open-ended, multiple-choice, true-false, completion, matching, rating scales, semantic differential, check lists, and so on.
4. *Mode of Expression for Items.* Verbal or non-verbal, oral or written.
5. *Student Abilities.* Types of thinking skills to be tested, such as memorization, recognition, reproduction, comprehension, analysis, and so on.
6. *Assumptions about Previous Knowledge, Experience, and Language Abilities of Subjects.* How much can it be assumed that those being tested will bring to the test situation as prior knowledge and what kinds of intellectual abilities and social backgrounds can they be assumed to have? At what reading level is written material to be cast?

7. *Group Size.* Are instruments to be administered by teachers, by trained testers, by older students, by parents, or by whom? What number of individuals will complete the instrument together? How detailed should the directions on the instrument be?

8. *Data Sources.* From whom will data be gathered: school administrators, teachers, students, observers, school records, etc.?

9. *Schedule of Use.* How many minutes do respondents have available for supplying information?

10. *Training Time Available.* How much time do instrument administrators have available for training in instrument use?

11. *Scoring Options.* What methods of scoring are preferable and are permissible?

12. *Data Coding.* How are items to be coded for later analysis?

Develop Items for Instruments

Items can be developed by specialists in the field test headquarters staff or by outside specialists brought in for the purpose. One or more outsiders can be used, depending on the timetable for creating the items. Item writers can be paid by the hour or they can be paid for piece work with payment made only for acceptable items and without regard to the amount of time taken to create them.

Item writers who are not familiar with the product being tested or with the purpose of the field test will need to be thoroughly trained in both before they can write satisfactory items.

Assemble Items Into Instruments

After the items have been produced, they should be classified according to the sources from which the data are to be gathered (school administrators, teachers, students, observers, school records, etc.).

Items should then be subdivided according to the time schedule on which the data must be collected (before the field test, periodically during the field test, at the conclusion of the field test, at some point after the field test has been concluded, and so on).

Once the items have been so classified, they can be compiled into instruments according to the specifications for each instrument to be administered to specified populations at specified times. Then all instruments can be assembled into a complete instrument package.

It is essential for field testers to remember that when the time comes for data analysis, the analysis must not be constrained in any way by the data source, the time at which the data were gathered, or the particular instruments with which the data were collected. It is not uncommon for data collected from a single source at a single time on a single questionnaire to be analyzed as an interrelated set even though it would make more sense to associate certain parts of the data set with data gathered from other persons at other times through other instruments. That is, the completion of items and the assembly of instruments should not be allowed to influence the pattern of analysis: the data collected by the instrument package should be disassembled and then reassembled according to the analysis requirements, regardless of source, time of data collection, or the instrument used to collect it.

Testing the Instruments

All instruments should be tested as a part of their development This is especially important in the case of field test instruments because they are likely to be used by a large number of people. Time must be specifically allowed in the overall schedule for testing

the instruments. The exigencies of field test circumstances and the considerable time required for instrument development have a tendency to absorb the time needed for instrument testing. Despite this natural tendency there is no complete substitute for the actual pilot use of an instrument to demonstrate whether the items mean to the respondent what they mean to the designer. Observing the instrument in use even by a few subjects on a few occasions can be extremely illuminating for the instrument developers and can lead to immediate, substantial improvements in the instrument.

The criteria which can be used in judging data-collection instruments include the following:

1. *Validity.* Does the instrument measure what it is intended to measure? Do the scores it yields correspond to the existence of the phenomena of interest in the real world?

2. *Reliability.* Does the instrument produce the same score in repeated uses if the phenomena being measured have not changed? Does it measure without fluctuations in the instrument itself?

3. *Objectivity.* Would two or more independent observers gather identical data when observing the same phenomenon? Does the instrument take measurements independently of the characteristics or skills of those who are using it? If not, can all users be made equally skilled?

4. *Utility.* Is the instrument easy to use? Does it require complex, hard-to-teach skills?

Data Collection

Collecting data includes both administering the instruments and shipping the completed documents back to the evaluation headquarters for processing. Instrument administration is a critical step in field testing. The conscientiousness and technical proficiency with which it is done determines the quality and the completeness of the data gathered. This in turn affects the final results.

While not as logistically complex as some other aspects of field test operation, careful attention to procedures used for returning the completed documents is essential in dealing with the cooperating field sites. Data can be lost at this stage through misplacement, misidentification, losses in transit, and so on. Thus the topic deserves careful thought and precise planning.

Administering Instruments

Instrument administration entails determining the personnel to be employed, the arrangements for their training, and the timetable to be used. These topics are discussed below.

Personnel. There are at least five groups of persons who can administer instruments. One or more of the following groups may be used, depending upon instruments employed, skills needed, confidentiality required, timing cost, and other factors. The advantages and disadvantages of using each are discussed briefly below.

School Staff. Paraprofessionals, classroom teachers, guidance counselors and other school specialists, and occasionally building administrators can administer data collection instruments. Their services are especially appropriate for instruments already familiar to them, such as standardized tests and simple questionnaires. And they are suitable where confidentiality of responses is not important. Moreover, using the school staff to administer the instruments will prove to be least disruptive to the school program and least expensive in most cases.

On the other hand, local personnel may introduce a favorable bias into the data if the instruments allow room for judgment, as in the case of observation guides which require the user to make some assessment of what he is observing. Thus administration by school personnel is the preferred method for simple or familiar instruments which are not subject to bias by the administrator and where the respondents have not been assured confidentiality. Conversely, it is not the best procedure for other kinds of instruments.

Some of the disadvantages of using the local staff can be avoided by selecting a small number of them to administer instruments and training them specifically for that purpose.

Project Staff. Sometimes a field test is run as part of a larger project which employs a full-time staff on site for assorted functions. These persons are generally employed by the developer. These persons sometimes can be used to administer instruments. Because of their small number, it can be economical to train them to use instruments unfamiliar to the school staff. Moreover, they may have more time available for the work than regular personnel. And they can be used to gather confidential data since they are somewhat detached from the regular school staff. Otherwise they have about the same advantages and disadvantages as regular school personnel.

Field Evaluation Staff. If activity at the site is extensive enough to justify the cost, field staff may be assigned to each site either part-time or full-time to instruct local school personnel in field test procedures, monitor their work, provide liaison with field test headquarters, and administer instruments which cannot be administered by the school staff. The field staff can be trained at reasonably low cost to administer unfamiliar or complex instruments and can of course be trusted with confidential data. Their limited number offers the added advantage of assuring reasonable comparability in the data gathered from class to class, school to school, and site to site.

On the other hand, using them to gather a large volume of data will prove far more expensive than using the school staff for the same purpose. Moreover, it may be impossible for them to collect the data when a large volume must be gathered simultaneously. For example, if all classroom teachers must be interviewed during the first week of school, the field evaluation staff may not be able to manage that number of interviews themselves.

Central Evaluation Staff. Some types of data are so significant, so sensitive, or require such skill to collect that only highly qualified members of the central evaluation staff should do the work. One example is collecting the views of the superintendent of the local school district about the effectiveness, utility, and acceptability of the product being tested. Another example is collecting the views of key community spokesmen of the same topic. Generally, of course, using the central evaluation staff to administer instruments is far too expensive. Thus they must usually spend their time in training others in how to do it.

Outside Personnel. A special cadre of outsiders can be employed and trained to administer instruments. Examples of such outsiders are local parents (for administering simple questionnaires to other parents, for instance) and graduate students in education, sociology, or other social science fields who are being trained in research and evaluation techniques and are familiar with the use of instruments and the need for objectivity. Occasionally a cadre of highly specialized personnel, such as experienced senior professors, need to be employed to administer instruments requiring professional judgment. An example would be the use of observation guides to evaluate complex student performances in situational tests.

There are obvious advantages in using outside personnel: They may be geographically close to field test sites, reducing or eliminating travel costs; they can be employed only for the hours or days needed, saving salary costs; they can all use the same instruments simultaneously, gathering a large volume of data at one critical point in time when required by the field test design.

Training in Use of Instruments. To reduce or eliminate the need for supplying training in administering instruments, thereby saving both time and money, field testers should use types of instruments that are already familiar to school personnel unless there is a compelling reason to use other types. If unfamiliar instruments must be used, they should be made self-instructional insofar as possible. If unfamiliar, complex instruments that are not actually self-instructional must be used, they should be designed so that a minimum of training is needed to use them. Other types of instruments requiring extensive and expensive training are likely to call for behaviors which the instrument administrators cannot be relied upon to use faithfully once they return to the field sites.

If training is necessary, it may be given centrally at evaluation headquarters, regionally, or within the individual sites. In general, centralized training allows the use of qualified members of the evaluation staff as trainers, is more likely to produce comparable behaviors in all the instrument administrators, is faster, and costs less. On the other hand, on-site training offers the advantage of being tailored to the specific needs of the trainees and the specific characteristics of the site. Training at regional locations tends to compromise the advantages of central training and site training and may prove to be the best choice.

Timetable. Data collection during field testing is likely to extend over a long period of time, especially if the product requries an extended period of use. Setting a careful timetable for instrument administration is important. Developing a PERT-like chart to examine the complete network of events can lead to such efficiencies as gathering data from students and teachers simultaneously and sequencing instrument administration so that a small, trained staff can handle most or all of the job. Data may be collected before, during, or after field testing.

Before Field Testing. After the sites have been selected but before the product testing actually begins, data on site characteristics can and should be collected. Gathering such data at this point makes more efficient use of the data-gathering staff and the data-processing staff and avoids burdening them with this responsibility after the field test begins. It also allows for early reports on site characteristics, which may be of interest to the sponsoring agency, the product developers, and outside audiences as well.

During Field Testing. This period covers the time when the product is at the site.

1. *Before Product Use.* Required pretests, measurements of teacher or pupil attitudes toward the product or the area of study before they begin using the product; status studies of the school program before the product has a chance to influence it. These are the kinds of data to be gathered before product use.

2. *During Product Use.* Day-to-day or week-to-week reactions to the product by students and teachers; daily work samples; weekly or monthly interim tests of pupil learning; teacher or student suggestions for modifying individual lessons or specific instructional materials such as individual films; opinion of school administrators about the utility, effectiveness, and acceptability of the product; early or interim community reaction to the product or to the field test procedures themselves; formal or informal monitoring reports about how the field test is proceeding. These are the kinds of data which the timetable must schedule for collection during product use.

3. *After Product Use.* Final tests of pupil achievement; overall teacher appraisals of the product and estimates of its effectiveness; overall evaluation by administrators; final student opinions; community reactions; summaries of time devoted to assorted product-using activities; inventories of product components available for re-use or for return to project headquarters; summary reports by site monitors. These are the kinds of data to be gathered after the product has been used.

After Field Testing. It maybe necessary at some point after the field test has been completed to gather additional data. For example, a full-year instructional program for graduating seniors perhaps cannot be finally evaluated until graduates have been in college or in their first jobs for several months. Final observations or interviews with the graduates or their employers may be scheduled to occur several months following graduation. (Strictly speaking, of course, the field test may be considered as still under way until all data have been collected, analyzed, and reported.)

Collecting and Returning Data

Once the instruments have been administered and the data recorded, the completed documents must be collected, inspected, properly labelled, assembled, packed and shipped to field test headquarters or to some other designated location.

Final Checking. Before completed documents are sent to the data processing location, it is essential that they be inspected for completeness and proper identification. It is extremely difficult and can be very expensive to do this after the documents have been sent away from the local site. While they are on site, it is often a simple matter to compare handwriting, look once more into the bottom of the box, check under the paper clips, make telephone calls to teachers, and otherwise do what is necessary to make sure that all documents are on hand and are properly labelled for shipment.

Shipment Timetable. The timetable for sending completed instruments to the data processing location is likely to be dictated by the data processing timetable itself. Documents may be shipped serially as they are completed or may be held for simultaneous shipment after all of them have been gathered. The first method is more likely to prevent data loss and preserve confidentiality but will probably be more expensive. The reverse is true for the second method.

Data Processing

Data processing consists of such activities as the following:

1. Labelling all documents to make sure that the data taken from them are properly identified henceforth. This should be done at the local test sites before the data is forwarded to a central data processing location.
2. Coding verbal labels into numerical form for ease of data manipulation during analysis.
3. Assembing documents into orderly sets to simplify overall organization.
4. Organizing data to interrelate various kinds of information, depending upon the analysis plan.
5. Scoring tests, questionnaires, rating sheets, observation guides, and other documents in which clusters of data are to be represented by single numbers.
6. Summarizing data a various levels of aggregation as required by the analysis plan. Typical levels of aggregation are classroom, school, school district, and type of school district.

7. Analyzing data to detect patterns, changes over time, differences between and among various groups, inter-relationships among variables, and so on.

8. Tabulating the data after analysis in displays needed by those who must interpret the findings and by those who must write the field test reports. Some of the tabulations will be designed for use solely by the field testing staff; others will be designed for inclusion in the field test reports.

9. Storing data following processing to make it conveniently available for further analysis should that be necessary in the future.

Any one or any combination of these processes can be handled either manually or by an array of data processing machinery ranging from keysort equipment through test scoring machines and desk calculators to high powered computers.

Any one or any combination of them can be performed centrally at field test headquarters or at some other designated location or they can be decentralized and performed at one or more of the field test sites.

Any one or any combination of them can be carried out serially as data are collected during the field test or can be delayed until the end when all data have been gathered.

The advantages and disadvantages of the various choices are outlined briefly below.

Methods

The choice between manual and machine methods is not automatic. A full-scale field test is likely to generate data requiring the use of both methods.

Manual Processing. Manual processing is sometimes desirable, sometimes necessary, and sometimes both. In general, manual methods are preferable to machine methods when the data have the following characteristics:

1. They are transitory, as in the case of monitoring reports containing information on the progress of the field test.

2. They cannot easily be reduced to numerical form, as in the case of reports describing the degree of cooperation exhibited by local school administrators in meeting field test requirements.

3. The volume is low, as in the case of demographic characteristics for school attendance areas supplied by building administrators.

4. They are non-repetitive, as in the case of data on school facilities, which can be gathered and processed once and once only.

5. They are not to be interconnected with other data, as in the case of local schedules for administering field test instruments.

6. They are not easily summarizable, as in the case of marginal notes made on copies of the product by classroom teachers giving their suggestions for revision.

7. They involve configurations of data which must be inspected and judged or scored as a set, as in the case of a pattern of responses given by a single respondent on more than one instrument.

8. They require a high degree of informed judgment, as in scoring essays or samples of artistic work.

9. There are no immediate deadlines for completing the process as the the case of data gathered continuously but not usable until all of it has been collected. This may be true, for example, of data on time elapsing during the use of various product segments.

10. The time and expense of setting up machine processing would approximate the cost of manual processing.

Machine Processing. The circumstances in which machine processing is indicated are essentially the opposite of those described above for manual processing. That is, the data are of permanent interest, they are easy to represent in numerical form, the information volume is high, the data are repetitive, they must be interconnected with other data, they can be readily summarized, scores can be assigned to parts rather than configurations, handling them does not require professional expertise, it is important to process them promptly, and the ratio of machine set-up time and cost to total processing time and cost is low.

In general, however, field testers will want to design instruments and lay out analysis plans that will allow the use of machine processing whenever feasible. The speed and economies to be gained – beginning with coding and assembling the data and ending with printing tabular displays for direct photographic reproduction and inclusion in the published reports can be enormous. This is especially true for field testing since it ordinarily entails handling a very large volume of data.

Location

As indicated above, any one or any combination of the many steps involved in data processing can be performed at a central location such as field test headquarters or at one or more of the field test sites. The factors to be considered in making the choice include the following.

Accuracy. Do the local sites have qualified personnel available to take the care necessary for accurate coding, scoring, analyzing, or whatever data processing steps are being considered? Inaccuracies in data handling – especially in preparing material for machine processing – can lead to costly delays at later stages. Unless accuracy can be guaranteed at the local sites, central processing is mandatory.

Consistency. Can it be assured that all local sites will use identical procedures in processing data? This is important, of course, in making the data comparable from site to site. To accomplish such consistency, training is usually necessary. One clear advantage in central data processing is consistency.

Time. Time can be saved by arranging for local processing, especially at the early stages when data are being labelled, coded, assembled, and organized. Scoring may or may not

be more easily handled at a local site, depending upon the caliber of local personnel available and the assurance that they will be accurate and consistent in their work. Ordinarily, it is time-saving to have the final stages of data processing – analysis, tabulation, and storage – performed at a central location since the need to interrelate data from the various sites virtually dictates central processing at that point.

Cost. The early stages of data processing can usually be handled by local personnel at local cost, resulting in a dollar saving for the field testers. In some cases, one or more local sites will have access to data processing machinery at lower costs than are available to the field testers themselves. In such cases there is obviously a cost advantage in having data processing done locally. Again, however, there is no advantage unless accuracy, consistency of data treatment, and timely completion of the work can be guaranteed.

Supervision. While supervision of data processing is necessary no matter where the work is located, additional supervision is needed if the work is decentralized. There is no other way to assure that the data are being properly handled.

If the work is decentralized, it is very important for the field testers to establish that local personnel have the skills necessary to do the job and are in sufficient control of their time that any data processing work they undertake will not be dislocated by irresistible competing demands from higher local officials.

Timing

Data processing may span a long period of time, particularly if data are being processed serially as they become available. Thus a complex scheduling problem faces the field tester in this respect, as in many others.

There are advantages in processing the data serially as it is being collected and there are compensating advantages in processing it simultaneously after all of it has been gathered.

In general, serial processing spread over a number of weeks or months involves fewer people, avoids delays in issuing reports at the end of the field test, and is more expensive. Conversely, saving all data processing until the field test has been completed usually requires more people to work during the period of processing, tends to delay the completion of field test reports because the heavy load of data often jams the data processing system (be it manual or mechanized) but ordinarily costs less than serial data processing. The differences are due in part to the fact that a smaller number of people can handle the work by spreading it out over a longer period of time when data are processed as they become available, reducing the need for training and putting the work in the hands of more people, and because spreading the work tends to eliminate heavy backlogs and allows time for correcting processing routines if they prove faulty when first used. The higher costs of serial processing arise from the fact that people are kept on the payroll for a longer period of time, which almost always raises costs, and because certain of the efficiencies available in batch processing are lost.

As in so many other aspects of field testing, there are advantages and disadvantages to each choice. One of the primary functions of planning is to explore alternatives and to make choices rationally with the best possible advance knowledge of the consequences.

CDC

The PSE staff offered the funding agency three choices of techniques for producing the final field trial report on each of the CDC units (25 reports in all): 1) an individually-written report, prepared separately by hand for each unit, 2) a semi-computerized report with computer-printed tables and some computer-printed

captions and commentary or 3) a fully computerized report with all tables and a narrative generated by computer. The fact that the first choice saved on computer programming time but cost more in writing time while the third choice did the reverse was explained when the alternatives were presented.

After consideration, the agency chose the middle alternative. The result was 25 reports, each consisting of standardized tables, about half printed by computer and about half typewritten, computer-generated captions and a narrative, and with longer sections of hand written narrative and interpretation. This arrangement greatly expedited report preparation and production.

Data Analysis

The plan for data analysis must correspond point-for-point with the purposes established for the field test. More specifically, the data analysis must be designed to answer the questions asked by all the decision-makers who will depend on the answers in whole or in part to guide their decisions about the product. As indicated earlier, those decision-makers will include the agencies sponsoring the field test, the developers of the product, those who must revise the product following field testing if necessary, prospective adopters of the product, and other audiences specified by the sponsoring agency when the field test was commissioned.

Inasmuch as the field test probably will have multiple purposes and inasmuch as the several decision-makers are very likely to have multiple questions, the data analysis will have to be multiple in character.

It is common for the study of data like those arising from a field test to be handled through widely known and readily available analytic routines. In the case of field testing however, it is best not to run the data through analyses of variance or multiple regression or canonical correlation or factor analytic procedures simply because they are customary and familiar (or because canned programs for them exist at the computer center). Field

testers need to demonstrate to themselves – or better still – to the audiences for the field test reports when those reports are first being designed – that the proposed analyses are exactly those needed to answer the questions the field test has been commissioned to answer.

In an earlier discussion it was noted that it is possible to gather more data than are needed to answer the questions addressed by the field test. Similarly, it is possible to perform more analyses than are necessary to answer those questions. Field testers are free, like any other investigators, to sift through their data looking for the unexpected. There is certainly no harm in that, if time and funds allow. But serendipity will not substitute for the careful derivation of the analysis plan from the questions that various audiences want the field test to answer.

The steps in developing the analysis plan include the following:

1. Review the purposes of the field test, which may include such objectives as:
    a. to show whether the newly-designed product performs as intended under normal or near-normal use.
    b. to show under what conditions the product does and does not perform as intended.
    c. to determine the costs of installing the product.
    d. to guide the revision of the product and/or the development of a substitute for it or supplementary products.
2. Re-examine the product to make certain that the degree to which it has accomplished all of its own objectives is being addressed in the analysis.
3. Review the criteria specified for evaluating the effectiveness of the product. These are likely to include:
    a. Product effectiveness

b. Product utility

c. Product acceptability

d. Effort required

e. Time required

f. Cost/benefit ratios

4. Review the data available on site characteristics, conditions of product use, attitudes toward the product, effect of the product on the school program itself and pupil performance, to make sure that independent and dependent variables are identified and that the analysis will relate them properly to each other.

5. Explicate how the full set of data is to be constructed, showing subdivisions, levels of aggregation, statistics (such as ratios) to be created by combining measures, and so on.

6. Detail the comparisons to be made, as between pretests and posttests, experimental and control groups, pupils of high ability and low ability, and so on.

7. Detail the interrelations to be established among the various measures, as between independent and dependent variables.

8. Specify non-statistical procedures to be used in analyzing qualitative data, such as that contained in monitoring reports or collected in informal feedback from school staffs.

9. Specify methods of statistical analysis to be applied to quantitative data, such as pupil performance test scores.

10. Identify tests of statistical significance to be employed in judging the sizes of differences obtained.

When the plan has been completed and the data begin to flow in to the processing center, it is wise to run small data samples through the analytic procedures to make sure that they are workable and yield interpretable results. Ideally, this will be done with simulated data well before actual data have to be dealt with so that the analytic procedures are ready to use when the real data arrive.

CDC

In developing the data analysis plan, the PSE staff argued that the field trial report should include not only raw findings but an interpretation of those findings including the answers to these summary questions:

1. How well was the unit learned?
2. What specific parts of the unit were learned?
3. Which children learned the unit?
4. What is the effect of teacher characteristics on pupil learning?

The PSE staff also argued that the report should include a set of suggestions for unit revision, based on the field trial findings.

RDO argued that the field trial reports would be just as useful to it – and quite possibly more useful – if they were limited to summaries of the data analysis, presented without interpretation, conclusions, or recommendations. RDO said, for example, that suggestions for unit revision ought not be based on the field trial reports alone, but should incorporate other ideas for unit revision which it had been gathering in a process parallel to the field trial. (As mentioned earlier, RDO conducted debriefings of some teachers who had taught the units, collected copies of the units, with marginal notes made by teachers during the field trial giving their suggestions for unit revisions, and reviewing what its editorial boards had said about the units when they were being readied for field trial.)

This debate represented a continuation of the long-standing conflict about whether the field trial should serve summative or formative purposes. The funding agency's decision, consistent with its series of earlier decisions, was to endorse the report format recommended by the PSE staff, since that would best serve the summative purposes which were uppermost in the agency's mind.

IV. REPORTING

Once the field test is under way, reports must be issued both during the test and after it has been completed.

During Field Testing

The sponsoring agency will undoubtedly want progress reports as the field test proceeds to make certain that the work is going according to schedule. These reports should be oral as well as written – oral for promptness and when an immediate reaction is needed and written so as to make a permanent record. The product developers, although they are somewhat less likely than the sponsoring agency to want progress reports, may be quite interested in how the product is being accepted and seems to be performing under field conditions. It is less likely that outside audiences will want progress reports and it is unlikely that the field test sponsor will want to pay for such reports to be produced, but it is possible. For example, if one of the purposes of field testing is to build an eventual market for the product, the sponsor may want the field testers to issue periodic reports to professional associations so that their journals can keep members informed in order to build and sustain interest in the product.

It is not uncommon to run into problems (or, more rarely, solutions) that require a change in field test procedures and/or funding. Keeping the sponsoring agency informed about progress as the field test proceeds will let it anticipate possible requests from the field testers for a change in procedures or in funding. This will let the agency give the field testers a realistic prediction of whether requests for changes will be approved. It is very important for field testers to anticipate how the sponsoring agency would react to change requests because unless the funds can be adjusted, [illegible] procedures undoubtedly

will have to be adjusted to meet emergencies. The earlier the field testers learn whether funding adjustments will be made by the sponsoring agency, the more time they have to adjust their procedures if necessary before going too far with the work. For example, if expenses run much higher than originally anticipated and it is necessary to cut costs, this may have to be done by reducing sample size. Once actual field work has been initiated, to cut sample size by aborting the effort in some of the cooperating local school districts will unquestionably cause negative reactions on the part of local personnel. To make such reductions before the field test actually gets under way is better for all parties.

After Field Testing

Once the field testing has been completed, reports must of course be issued to describe the success or failure of the product in achieving its objectives. Such reports will have to describe the conditions under which the product was tested and will have to detail the evaluation procedures carried out. The report should also include information on "side effects", such as product acceptance by the users, interest generated in acquiring additional products of the same type, disruption in the school program, public reaction to the use of the product, and so on.

In developing progress reports during the field test as well as in developing product reports after the field test, careful consideration must be given to the audience for the reports, and to the content, format, timing, and distribution of the reports. Decisions about these matters must rest on the purposes for conducting the field test.

Audiences

For progress reports as well as for product reports, the intended audiences and their information needs must be carefully considered. This is not the final step in field testing;

it is the first step. That is, well before the field test is designed, evaluators must identify the key recipients of the reports, identify the decisions that they must make based on field test findings, and list the questions they need to have answered in order to make those decisions. The field test must then be designed to arrive at such answers and the reports written to deliver those answers. Obviously, the purposes of the field tests will help determine the audiences – and, of course, vice versa.

The field testers must determine the audiences jointly with the field test sponsors and with the product developers. The sponsors and developers will, of course, be the primary audiences to be served. In addition to themselves, however, they may specify potential adopters of the product being field tested, developers of products similar to the one being tested, developers at large who may be able to profit from the findings, other field testers who may be able to benefit from the methodology used in the field tests, and so on.

It is not satisfactory for the field testers to guess about audience needs or to use second-hand information. Each audience has to be dealt with directly. Field testers will often be surprised about what various audiences want to know. It may be more or less than what the field testers would imagine and it may be quite different information than they anticipated. While there will be overlap among audiences in their information requirements, each is likely to want unique information as well. One good way for identifying the content, format, and timing of reports to various audiences is to develop samples of alternative reports and to submit these to the several audiences so that they can select what they want. Nothing draws sharp reactions and specific guidance so well as an actual sample of what the field test proposes to give the audience. Money spent early to pinpoint audience preferences for the content, format, and timing of reports has far greater value than money spent later to change unsatisfactory reports.

CDC

The PSE staff did not make any systematic survey of possible audiences for the field trial results during the months of debate about what should be included in them. The staff talked only to RDO and the funding agency.

Later, after all 25 reports were printed, in conducting a federally-sponsored study of the evaluation data needed by decision-makers in educational programs, the PSE staff identified a number of audiences who wanted information about the field trial – but not in the form of 25 long reports. The following federal government officials all said that they would like a summary description of the performance of the 25 CDC units – not a report for each unit, but an overall report for the units as a group:

HEW Staff

NIE Administrative Staff

Office of Education Administrative Staff

RDO Management Staff

If that interest had been known at the inception of the field trials, it would have been relatively easy to design and prepare such a summary report.

Content

As indicated above, the content of the field test progress reports and final reports should be dictated by the questions which various classes of decision-makers (audiences) need to have answered. The content may not be the same for each audience, but the range of information needed is indicated below.

Progress Reports

Progress reports should contain information about the extent of on-time performance, discrepancy between procedures originally planned and procedures actually used, problems encountered, solutions adopted or proposed solutions for which approval is being sought from the field test sponsor, apparent degree of product acceptance by various users, and early impressions of product outcomes.

Product Reports

The final field test reports should include information such as the following, depending upon the intended audience.

Developmental History of the Product. The reader of the field test report may be interested in how, when, and by whom the product itself was developed. If a complete description is available elsewhere, a brief summary and a reference to the fuller description will be sufficient. Otherwise a longer description of the developmental history should be written.

Description of the Product. The field test report will be meaningful only if the reader is thoroughly familiar with the product being evaluated. If a longer description is available elsewhere or if it can be assumed that the reader of the report will have direct access to the product itself (this depends on the audience as well as on the complexity, expense, and ready availability of the product) a brief description will be essential. But if a full product is not easily available to a particular audience, a longer description of the product will be needed in the field test report.

Objectives of the Product. The behaviors that the product is intended to produce in its users should be described in the field test report. These are, of course, the objectives against which product performance is to be reported.

Description of Agencies Developing the Product, Sponsoring the Field Test, and Conducting the Field Test. Most audiences will want to know who developed the product, who initiated and is paying for the field test, and who is carrying it out. This information is important in case the reader wants to make further inquiries into the nature of the

product and its availability, the results of the field test, or the procedures used in conducting the field test.

Purposes of the Field Test. Inasmuch as the character of the report is dictated by the purposes of the field test, the purposes should be indicated.

Procedures Used for Product Evaluation. The report should contain information about how the sites were selected; what data were collected; what data-collection instruments were adopted, modified, or developed for the field tests; the size and characteristics of the sample drawn; the data-collection procedures; techniques of data processing; and methods of data analysis. These kinds of information will be of more or less interest to various audiences.

Conditions of Product Use. Most audiences will be quite interested in the conditions under which the product was tested, since these are very likely to influence the outcomes. The conditions found to be critical to product success will, of course, be of special interest to all audiences.

Attitudes Toward the Product. The attitudes of the administrators, teachers, students, parents and any others who came into contact with the product directly or indirectly during the field test will be of interest to most audiences, since these attitudes will tend to predict eventual acceptance and diffusion of the product.

Product Effectiveness. The extent to which the product achieved its objectives with various populations is, of course, the most significant information in the field test report. All audiences will be interested in these findings.

Effects of Other Factors on Product Effectiveness. Apart from the conditions under which the product was used and the kinds of populations for whom it proved effective, any other variables associated with its performance would be reported. For example, the product may have worked better in some kinds of schools or school districts than in others. The relation of site characteristics to product performance should be indicated.

Cost/Benefit Ratios. If comparative information is available about the cost and effectiveness of other products having objectives identical to those of the product field tested, the relative cost advantage (or disadvantage, if that is the case) of the product under test should be reported. (The actual cost of installing the product and maintaining it in use will have been reported earlier, under *Description of the Product*).

Practicality of the Product. Field testing the product under natural conditions will yield information on ease of use, access to necessary facilities and equipment, and operational difficulties. This information will be quite important to prospective users of the product as well as to the product revisers who may want to change it to increase its utility.

Recommendations. Recommendations may be addressed to the sponsoring agency (for example, should further investment be made in developing additional products of the same type?) to product developers (for example, should additional products have somewhat different characteristics to improve their performance or increase their utility?) to product revisers (how should the product be changed to improve its utility and effectiveness?) to prospective adopters (how can the product be adapted to local circumstances?) and so on. Obviously, recommendations will need to be changed according to the action options open to the audience being addressed.

Appendices. Detailed statistical tables, technical descriptions of methods of data analysis, copies of evaluation instruments, and similar information too detailed for inclusion in the main text should be placed in the appendices.

Format

Report formats can differ radically depending upon their audiences. Reports can and will be transmitted orally when the occasion requires (oral reports are likely to be addressed to the sponsoring agency and perhaps to the product developers) or as written narratives or in the form of tables, graphs, and charts or in assorted audio-visual forms. Some audiences may want raw information presented first, with any interpretations, conclusions, and recommendations presented later and separately. Other audiences may want interpretations conclusions, and recommendations presented first, with the raw information being presented in a supporting appendix. An audience may want one without the other being presented at all.

Some audiences will want full technical reports, with heavy emphasis on field test methodology; other audiences will want popularized journalistic reports completely free of technical detail.

As explained earlier, the best time to get advice on formats is before the reports are designed and the best means is by presenting samples to the various audiences to be served. There are definite budget implications growing out of format preference.

For example, audio-visual presentations may be exactly right for certain audiences but they may be expensive to produce. Unless such reports are anticipated and planned from the very beginning, the field test budget is not likely to be sufficient for them.

Timing

When reports are first being planned, the several audiences should be interviewed to find their decision timetables. These timetables will dictate the timing of the field test reports.

In general, progress reports will be needed as promptly as possible since they are intended chiefly to identify problems and initiate corrective action. In contrast, product reports cannot be generated until after the field test has been completed. Even at this point, however, there is likely to be a time sequence in which various audiences should be served. Here is a possible sequence:

1. those who must revise the product,
2. the field test sponsors,
3. the product developers,
4. prospective users,
5. other field testers interested in the performance of various types of products.
6. other product developers interested in the performance of various types of products.

The field testers will find the scheduling and production of reports to be easier and more economical if they can work out the sequence so that the same report writers can work serially; otherwise they will have to employ, orient, and train a large group of writers to produce a complete set of reports simultaneously.

> CDC
> The PSE staff scheduled the production of the field trial reports serially, on the assumption that the CDC units would be tried out serially and revised serially. However, the accumulating lateness of the units resulted in the field trial data arriving at the field trial headquarters in late spring and early summer, giving the project staff an indigestible mass of data and report writing. Because the plan for data analysis and report writing had been carefully designed with expert personnel scheduled to work serially, producing four reports a month (a satisfactory schedule had the work begun in the fall of the field trial year as originally projected) it proved impossible to accelerate this schedule. The result was that it took months after the final CDC unit was tried out in June, to complete the final field trial report for that unit.

Distribution

Field test reports in written or audio-visual form will need to be presented or distributed to all audiences being served.

The field testers may be called upon to make platform presentations of their findings. On such occasions they will probably want to be equipped with handouts, overhead transparencies, slides, wall charts, tape recordings, or other supporting devices. Simultaneously, the staff of the sponsoring agency or the product developers may have occasions to make platform presentations and will need such supporting devices. The field testers may need to create and supply such devices to any audience needing them.

Multiple copies of field test reports are likely to be needed by most audiences. Determining the full number of copies before the first ones are reproduced will eliminate the necessity of reprinting and thereby reduce costs. Typing copy in camera-ready form and in formats suitable for reproduction will also lessen costs and accelerate production by allowing the use of a photo-offset printing process.

Reports may be needed by field test sponsors, product developers, product revisers (if other than the original developers), potential product adopters, other developers, other field testers, and the general research and development community. Arranging to place reports in ERIC (Educational Resources Information Clearinghouse), other information banks, and libraries can guarantee the permanent availability of the reports to interested audiences after the initial supply of reports has been exhausted.

Additional copies of papers in this series are available for a fee of $3.00. Previous papers in this series are listed below. Please direct your inquiries to:

Mary Anne Bunda, Series Editor
The Evaluation Center
College of Education
Western Michigan University
Kalamazoo, MI 49008

1. *Evaluation Bibliography* edited by Mary Anne Bunda

2. *Hints for Planning and Conducting a Survey and a Bibliography of Survey Methods* by Randy E. Demaline and D. William Quinn

3. *Field Testing Educational Products* by Henry M. Brickell and Carol B. Aslanian